RADC-TR-66-6
Final Report

DISCRIMINANT ANALYSIS FOR CONTENT CLASSIFICATION

John H. Williams, Jr.
International Business Machines Corporation

TECHNICAL REPORT NO. RADC-TR-66-6
February 1966

Distribution of this document is unlimited

Information Processing Branch
Rome Air Development Center
Research and Technology Division
Air Force Systems Command
Griffiss Air Force Base, New York

DISCRIMINANT ANALYSIS FOR CONTENT CLASSIFICATION

John H. Williams, Jr.
International Business Machines Corporation

Distribution of this document is unlimited

AFLC, GAFB, N.Y., 29 Mar 66-124

FOREWORD

This report describes a study performed by the Federal Systems Division, IBM Corporation, 7220 Wisco sin Avenue, Bethesda, Maryland, under Contract AF 30(602)-3563 to the Rome Air Development Center, Air Force Systems Command, Griffiss AFB, New York.. The study, performed during the period November 23, 1964 to November 22, 1965, considered the principal parameters affecting content analysis of documents by a technique based upon multiple discriminant functions.

The author is indebted to Mr. Robert Ruberti of the Rome Air Development Center for his valuable suggestions, encouragement and cooperation during the course of the study; to Mr. Gary Johnson of the IBM Corporation for his assistance in programming and performing the experiments; and to Mr. Robert Phillips of the IBM Corporation for his consultation in devising the hypotheses and experiments.

This report has been reviewed and is approved.

Approved:
FRANK J. TOMAINI
Chief, Information Processing Br
Intel and Info Processing Division

Approved:
ROBERT J. QUINN, JR., Colonel, USAF
Chief, Intel and Info Processing Div

FOR THE COMMANDER:
IRVING J. GABELMAN
Chief, Advanced Studies Group

ABSTRACT

A series of experiments was performed to investigate the effectiveness and utility of automatically classifying documents through the use of multiple discriminant functions. Classification is accomplished by computing the distance from the mean vector of each category to the vector of observed frequencies of a document and assigning the document to the category having the highest probability. Data concerning the effect of the principal classification parameters on classification performance is reported, based on a data base of approximately 2700 abstracts from the solid state physics field. The parameters studied were the number of sample documents required to define a category, the length of documents, the interrelationship of the number of sample documents and their lengths, the relation of the number of word types in a document to the number of categories assigned to it, levels in a structure, homogeneity of categories, and performance measures. A higher performance level was obtained when samples of 140 documents were used to define each category than with samples of 35 and 70 documents. Classification results obtained on independent test sets of documents ranged from 73 to 92 percent. The test sets contained 419 and 1333 documents. Results are also reported in terms of Swets' effectiveness measure and Cleverdon's ratios of relevance, recall and precision.

EVALUATION

The purpose of this work was to identify parameters which affect the accuracy of an automatic document classification technique based on the use of multiple discriminant functions. This study was necessary prior to testing the technique in an operational environment.

The most significant parameter found was the number of sample documents used to define each category. By increasing the number of sample documents, improvements were realized in both accuracy and the set of discriminating word types selected to represent each category. In addition, document length was found to have no adverse affect on accuracy. Arrival rate of new word types approached a reasonably small number and accuracy improved at lower levels in the classification structure.

As a result of this effort, this technique will be tested in a user environment where improvements from user suggestions and reactions should result in an operational classification technique.

Robert N. Ruberti

ROBERT N. RUBERTI
Project Engineer
EMIIH

CONTENTS

CONTENTS (Continued)

ILLUSTRATIONS

TABLES

TABLES (Continued)

Section I

INTRODUCTION

With the increasing use of optical character recognition equipment, direct teletype input and other forms of high-speed, high-volume input devices for computers, the processing of narrative documents is becoming more practical. Methods are required for screening, writing, and disseminating such documents; for indexing and abstracting them; and for extracting important information from them. While techniques for performing such functions automatically on formatted or semi-formatted documents have been in use for some time, the low density of information and wide variety of ways of expressing it have made the development of methods for processing narrative text more difficult.

This study was undertaken as part of a continuing effort to develop a useful and reliable technique for automatic content analysis of documents containing primarily narrative text. Content analysis is a key element of the screening, routing, disseminating, classifying and indexing functions. In general, the present work is directed toward automatically assigning a label or labels from a predetermined, finite set of possible labels to a given sample of text. The particular application can be, for example, assigning index terms, choosing nodes in a classification structure, determining whether a document should be sent to a certain analyst.

Several investigators have proposed techniques for these purposes using a variety of approaches. One particularly effective approach is the application of the statistical technique of multiple discriminant functions. This technique was first investigated in IBM's IRAD Task #0274 and is described in the task final report(1). A subsequent series of experiments was reported in a technical paper(2). Following two more successful sets of experiments, the present study was undertaken to determine the principal parameters affecting content analysis of documents as performed by the above technique. The technique was held constant while the effects of the various parameters involved in its use were studied. The data base on which the experiments were made consisted of a set of approximately 2700 abstracts of documents in the field of solid state physics. The results of the study lead both to effective application of the technique and to suggested ways of improving its performance.

Section II

OBJECTIVES OF THE CONTRACT EFFORT

The objectives of this contract effort were to determine the principal parameters affecting classification of documents by an automatic technique employing multiple discriminant functions, and to investigate their effect on the utility and effectiveness of the classification technique. The parameters to be studied were:

a. Number of sample documents used to define a category

b. Length of documents measured in terms of the number of word tokens and the number of sentences in it

c. Interrelationship between the number of documents and their lengths used to define a category

d. Relation of the number of word types in a document to the number of concepts (categories) assigned to that document

e. Relation between the level in a structure and classification performance

f. Homogeneity of a category

g. Metric used to measure the effect of the parameters on classification performance.

The technique requires a sample of documents known to belong to each category to compute coefficients for the classification decision. As the number of sample documents is increased, a better estimate of the discriminant coefficients can be expected. However, as the number of sample documents increases the number of different word types observed also increases. The two problems are to find a subset of types and a reasonable number of documents from which reliable estimates of means and dispersion can be computed. A sample is required for each category so that if a large number of documents were required for reliable estimates, the technique might become uneconomical. One would expect that fewer long documents are required to obtain reliable estimates; but since long documents are apt to add more new types to the word population, the need to study the interrelationship of the length of documents existe l.

Overall classification performance may be impaired by the nature of certain categories. These categories may be less homogeneous than others. Lack of homogeneity in a category may be due to the definition of its subject, the sample documents representing it, or the subset of words selected to represent it. A method of recognizing lack of homogeneity and its effect on classification performance needs to be determined.

Some classification techniques perform well when distinguishing between general subjects, but need further study to assess their performance at more detailed subject levels. Rather than classify documents on broad subjects such as chemistry, physics and biology, experiments were conducted at detailed levels within the field of solid state devices, distinguishing between the application, metallurgical, chemical and physical properties of devices. Further experiments were performed to distinguish between specific magnetic properties.

In operational situations many documents are classified into more than one category. A classification technique must also be able to adequately classify this type of document. Investigations are required to deter nine if this class of documents contains statistical properties different from documents belonging to only one category.

Since some of the parameters are interrelated, their individual effect in the multivariate model cannot be measured easily. A metric is needed to measure the contribution of each parameter to classification performance. As in all parameter studies the metric must be invariant with respect to the data set, otherwise comparisons between various applications would be impossible.

The objective of this set of tasks was to provide an understanding of the capabilities and limitations of a statistical classification technique. Isolation, definition and measurement of classification parameters is an important step that must be accomplished prior to prototype testing of any new technique. Once the ranges and effects of the classification parameters are established, applicability to specific systems, environments, or needs can then be evaluated.

Section III

CLASSIFICATION TECHNIQUE

This section contains an operational description of the classification procedure, a brief discussion of the multiple discriminant functions used, and a short summary of the computer programs used to implement the technique experimentally.

OPERATIONAL DESCRIPTION

The problems encountered in a classification technique are: selection of the variables on which the classification decision is to be based, computation of the classification statistics, determination of the decision rule, and analysis of effectiveness.

A user starts with a set of documents and decides on a group of subjects that interest him. He then constructs a tree showing the classification structure desired for the total collection of documents under consideration. Each node of the tree represents a category, and all nodes or branches emanating from that node are its subcategories. A category number is assigned to each node. Each document from a sample of the total collection is manually assigned to one or more nodes of the tree.

The variables used in the classification process are frequency statistics on word types from each category. Previous experiments indicate that all word types do not need to be retained for the classification process, and computationally it would be impractical to do so. Ideally, words selected to represent the categories should occur in one and only one category. However, there are usually only a few words in any data base that occur in one and only one category, and these words do not necessarily occur in every document. Therefore, a statistic is needed to identify words approximating this condition. Such words should have a small within-category variance and a large among-category variance; i.e., their frequency should be close to the mean for one category but far from the mean for all other categories. The F ratio of among-category category variance and pooled within-category variance meets this condition and has been used as a word selection statistic. The F ratio is also similar to the multivariate maximizing condition of discriminant analysis; therefore, words having a high value of F should also have high multiple discriminant coefficients.

An F ratio is computed for each word type after its frequency of occurrence in each sample document is counted and the mean frequencies, pooled within-category variances and among-category variances are calculated. A subset of words, called the discriminating word set, is selected for retention for computation of more precise weighting coefficients for use in the classification equation. Words selected for retention must not only have a high F ratio but must also have a high likelihood of occurring in many documents. (Otherwise, too many words would have to be retained to achieve an adequate level of coverage in each document.)

Finally, to represent each category in this experiment, s words are selected, where $s = 50/q$; q is the number of categories, and s is an integer formed by truncating the remainder of the quotient. The words in descending order of F are examined and for each category the first s words having a mean greater than 0.2 are selected. When this procedure is completed, $n = qs \leq 50$ words will have been selected. In the classification experiments 16 words were selected for each of the three categories. (A program limitation currently precludes the selection of more than 50 types.)

Having defined a multidimensional space with 48 words, the sample documents are used again to compute means and dispersion of each category and or the weighting coefficients of each word for the multiple discriminant functions. One set of multiple discriminant functions is computed for each node (i.e., group of categories) in a structure. The weighting coefficients for each word are determined by the context in which it is used. The center (centroid) of each category is represented by a vector of mean frequencies derived from sample documents known to belong to that category.

A new document is classified by counting the frequencies of the words occurring in it and comparing this vector of observed frequencies with the mean vector of every category. The probability of membership in each category is computed and the document is assigned to the category having the highest probability. For applications in which assignment to one category is not desirable, the probabilities may be stored for future retrieval requests. These probabilities may be considered relevance values, and responses to queries may be ranked in descending order of the probabilities.

MULTIPLE DISCRIMINANT FUNCTIONS

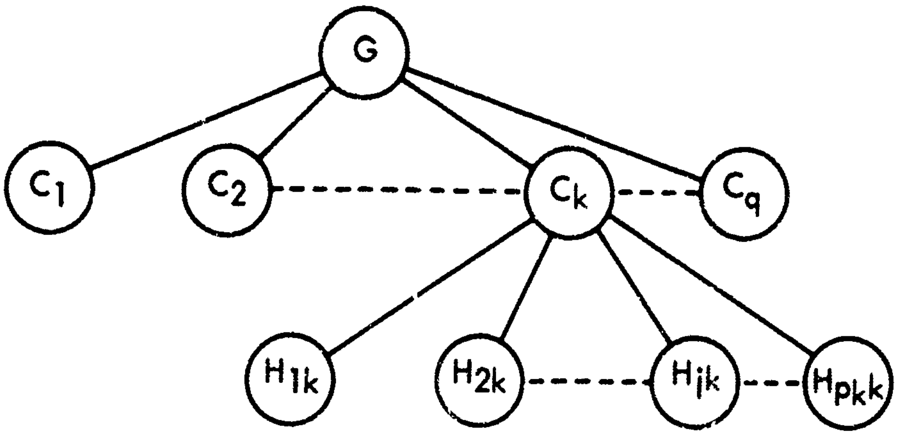
Many statistical techniques exist for the classification of a random observation into one of two populations. However, not until recently have techniques been developed for classifying observations into many categories. A survey of the techniques has indicated that multiple discriminant functions appear to be the best statistical technique for document classification. They are functions that contain weighting coefficients which indicate the discrimination ability of each variable. They also provide an advantage over linear discriminant functions since pairwise comparisons between two categories at a time are unnecessary. The relative probability of membership in each category can be obtained with a single statistic for each category.

Hodges[3], and Tatsuoka and Tiedman[4] have provided excellent surveys of classification and discrimination techniques. The original work in discriminant analysis was performed by Fisher [5] and which Barnard[6] applied to a problem dealing with the classification of skulls into archeological time periods. From a representative sample of each skull, the average value of each measurement was computed for each period. A new skull could then be classified by comparing differences in the observed value with the expected value of each of the seven measurements. Rather than sum these differences with equal weight, Fisher devised a technique for computing a coefficient for each difference. The desired set of coefficients would be the one that gave greater weight to those measurements which provided greater discrimination among the time periods. This set of coefficients yields a linear discriminant function. The optimum set of coefficients is found by maximizing the ratio of the among-category sum-of-squares of this function to its pooled within-categories sum-of-squares.

Further research has been performed in discriminant analysis by Rao[7] and Bryan[8]. Rao gives the necessary theoretical discussion and proofs for discriminant analysis.

The extension from the two-category to the q-category problem has been accomplished by Byran. He provided a method of obtaining more than one linear discriminant function and showed that a set of multiple discriminant functions can be obtained that exhaust all the information available concerning the separation of categories.

Prior to the computation of the discriminant function, a set of variables (words) must be selected. Since classification is performed sequentially, one level at a time, from the highest to the lowest level of the classification tree, a general notation can be expressed in terms of one level, as follows:



Let the group G consist of q categories, denoted by C_k $(k=1,2,\ldots,q)$. Each category is represented by p_k reference documents, denoted by $H_{jk}(j=1,\ldots,p_k)$. All reference documents in the group together contain n word types, say $Z_1, Z_2, \ldots, Z_n$. Let X_{ijk} represent the frequency of the i^{th} word type in the j^{th} document in the k^{th} category.

First, the mean number of occurrences per document of the i^{th} type in the k^{th} category is computed:

$$\overline{X}_{i.k} = \frac{1}{p_k} \sum_{j=1}^{p_k} X_{ijk} \qquad (1)$$

and from this the pooled within-category sums of squares are computed:

$$W_i = \sum_{k=1}^{q} \sum_{j=1}^{p_k} \left(X_{ijk} - \overline{X}_{i.k}\right)^2 \qquad (2)$$

Next, the mean number of occurrences per document of the ith type in the entire group is computed:

$$\bar{X}_{i..} = \frac{1}{\sum_{k=1}^{q} p_k} \left[\sum_{k=1}^{q} \sum_{j=1}^{p} X_{ijk} \right] \qquad (3)$$

and from this the among-category sums of squares are computed:

$$A_i = \sum_{k=1}^{q} \left(\bar{X}_{i.k} - \bar{X}_{i..} \right)^2 \qquad (4)$$

Finally, a discriminant ratio is computed for each word type:

$$F_i = \frac{A_i}{q-1} \div \frac{W_i}{\sum_{k=1}^{q} (p_k - 1)} \qquad (5)$$

$$= \frac{A_i \sum_{k=1}^{q} (p_k - 1)}{(q-1)\, W_i}$$

Having defined a multidimensional space with 48 words, the coefficients of the discriminant function can be computed from a sample of documents previously assigned to each category. Let the matrix A consist of the among-group cross-products of deviations of category means from the group mean weighted by category sizes:

$$a_{ih} = \sum_{k=1}^{q} p_k \left(\bar{X}_{i.k} - \bar{X}_{i..}\right) \left(\bar{X}_{h.k} - \bar{X}_{h..}\right) \qquad (6)$$

and the matrix W consist of the pooled within-group deviation cross-products:

$$w_{ih} = \sum_{k=1}^{q} \sum_{j=1}^{p_k} \left(X_{ijk} - \bar{X}_{i.k}\right) \left(X_{hjk} - \bar{X}_{h.k}\right) \qquad (7)$$

where χ $(i,h=1,2,\ldots n)$ is now the frequency of only the discriminating subset of n words. Bryan has shown that the condition for maximizing the ratio of the among-category sums-of-squares to the pooled within-category sums-of-squares is satisfied by solving the determinantal equation:

$$\left|W^{-1} A - \lambda I\right| = 0 \qquad (8)$$

where I is the identity matrix, W and A are as defined previously, and λ is any one of the m eigenvalues to be determined. The eigenvector corresponding to λ provides the set of coefficients for a discriminant function which transforms the n individual measurements into a single value or discriminant score. The linear combinations corresponding to the resulting eigenvectors have the following property: the first linear combination, corresponding to the largest eigenvalue λ_1, maximizes the discriminant criterion in the sense that one is discriminating between two categories; the second linear combination, corresponding to the second largest eigenvalue λ_2, maximizes the ratio of the residual among-category sums-of-squares to the residual within-category sums-of-squares after the effect of the first has been removed; and so forth.

Furthermore, the number of eigenvectors of the determinantal equation such that $\lambda \neq 0$ is at most equal to the smaller of the two numbers q-1 and n. These eigenvectors are the basis for multiple discriminant functions (MDFs) and exhaust the total discriminative power ot the variables relevant to category separation.

The MDFs are a powerful tool in that they preserve the information given by the variables relevant to group separation, and yet allow one to classify in an m-dimensional reduced space, where $m = \min(q-1, n)$. Rather than classify in the original n-variable space, the eigenvectors are used to transform the mean vector, M_{ik} and dispersion matrix, D_{ihk} for the kth category to the classification or reduced space, having m dimensions, where classification can be performed more economically. No loss of information is incurred by the reduction of 48 dimensions to 3 dimensions. The m eigenvectors of the transformation matrix, V_{if} are

$$\underset{(n,m)}{V_{if}} = \begin{bmatrix} V_{11} & V_{21} \cdots & V_{m1} \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ V_{1n} & V_{2n} & V_{mn} \end{bmatrix} \tag{9}$$

The transformation equations are:

$$\begin{array}{ccccc} \hat{M}_{fk} & = & V^{T}_{fi} & M_{ik} \\ (m,q) & & (m,n) & (n,q) \end{array} \tag{10}$$

$$\begin{array}{cccccc} \hat{D}_{fgk} & = & V^{T}_{fi} & D_{ihk} & V_{if} \\ (m,m,k) & & (m,n) & (n,n,k) & (n,m) \end{array} \tag{11}$$

where $f, g = 1, 2, \ldots m$ and the line below each equation indicates the dimensions of the matrices. The two matrices $\hat{M}_{fk}$ and $\hat{D}_{fgk}$ thus define the category C_k in terms of its centroid and dispersion in the discriminant space.

New documents are classified by counting all its words and retaining only the frequencies of the n discriminating words. These frequencies X_{ij} $(i = 1, 2, \ldots, n)$ for the jth document are transformed to m-dimension space by

$$\underset{(m,p)}{\hat{X}_{fj}} = \underset{(m,n)}{V^{T}_{fi}} \; \underset{(n,p)}{X_{ij}} \tag{12}$$

X_{fj} represents the fth coordinate of the jth document in the classification space. Classification is accomplished by computing the distance the jth document is from each of the q categories and computing the relative probability of membership in the rth category by

$$\underset{(p,q)}{\chi^{2}_{jk}} = \underset{(m,p)}{(\hat{X}_{fj}} - \underset{(m,q)}{\hat{M}_{fk})^{T}} \; \underset{(m,m,k)}{\hat{D}_{fgk}^{-1}} \; \underset{(m,p)}{(\hat{X}_{fj}} - \underset{(m,g)}{\hat{M}_{fk})} \tag{13}$$

$$\hat{\pi}_{jr} (C_r | X_j) = \frac{\dfrac{p_r}{|\hat{D}_r|^{1/2}} \exp(-\frac{1}{2} \hat{\chi}^{2}_{jr})}{\displaystyle\sum_{k=1}^{q} \frac{p_k}{|\hat{D}_k|^{1/2}} \exp(-\frac{1}{2} \hat{\chi}^{2}_{jk})} \tag{14}$$

where p_k is the a priori probability of membership in category k, and $k = 1,2,\ldots,r,\ \ldots,q$.

If the distribution of the categories coordinates are normally distributed about the category centroid, then the square of the distance from the centroid has a chi-square distribution. The current decision rule assigns the jth document to the category for which $\hat{\pi}_{jr}$ is the highest.

PROGRAM DESCRIPTION

A set of experimental computer programs has been written, tested and used to perform document classification using multiple discriminant functions.

Figure 1 shows the overall logic flow of the classification programs. The Frequency Tape Generator selects documents from a master document tape that are required for a particular experiment, and extracts the frequencies of a given set of words from each selected document.

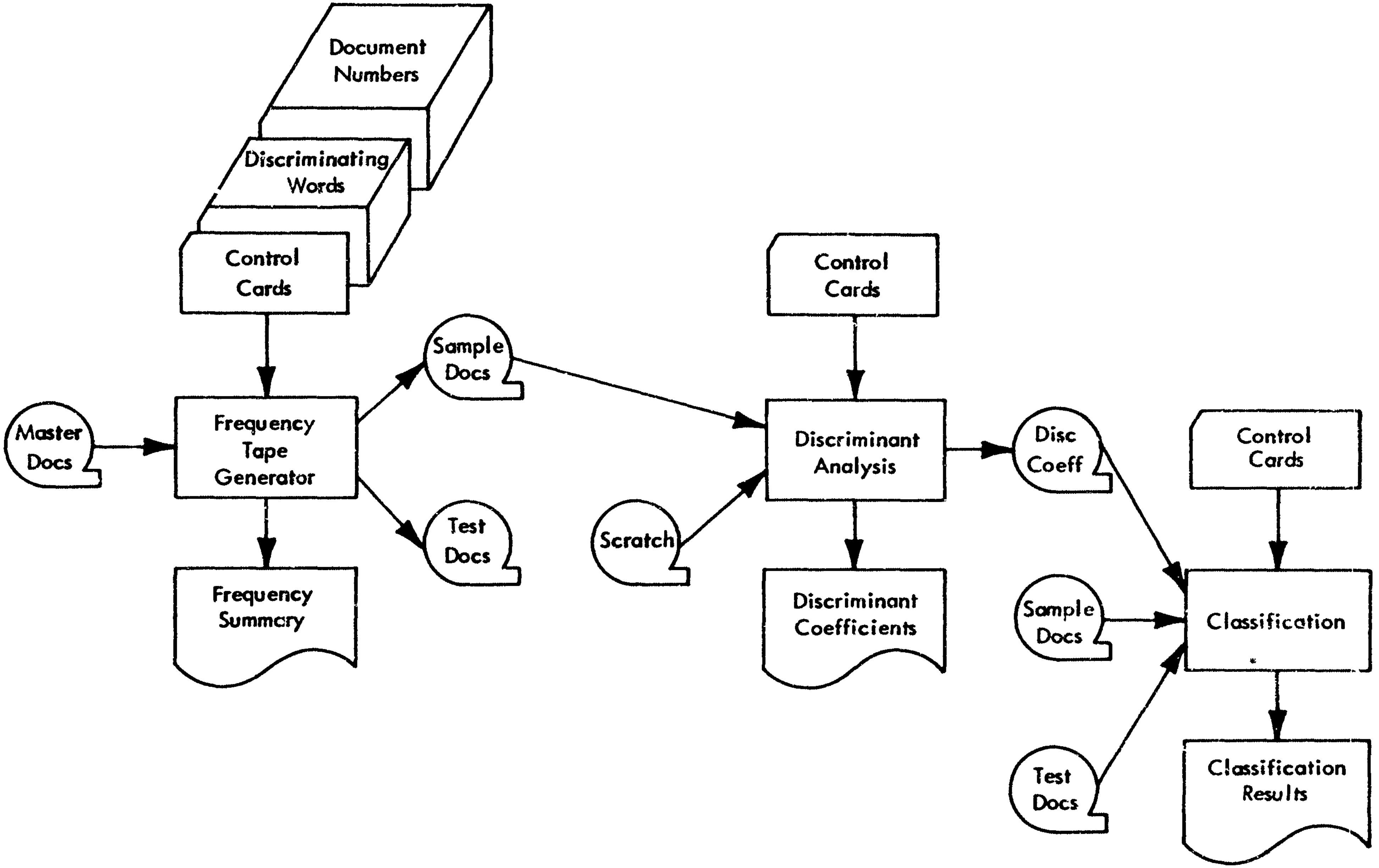


Figure 1. Discriminant Analysis Logic Diagram

The primary output is a binary document tape containing the document identification and a list of up to 50 words and their frequencies. The program will output either raw or relative frequencies depending on user need. A secondary output is a listing of the words and their frequencies and summary information concerning the distribution of words across the documents.

The discriminant analysis program computes the means and dispersion for each category, and the discriminant coefficients; it then transforms the centroids of each category to the classification space. The program will process up to 50 words, 10 categories and any number of documents.

The program reads a sample document tape output from the Frequency Tape Generator program. It outputs the discriminant coefficients on a tape for the classification program. The program outputs for off-line listing: the mean frequency and its standard deviation of each word in each category and the total group; the total correlation matrix; the eigenvalues and eigenvectors (discriminant coefficients); and the centroids and dispersions of each category in the original space and the reduced space. An option exists for a more detailed output of many intermediate matrices.

The classification program classifies documents and outputs summary information concerning the classification parameters. Each document is classified by transforming the frequencies of its discriminating words to the classification space. In the classification space the distance from the centroid of each category to the document is computed. The document is classified into the nearest category.

The program also outputs tables of the number of correctly and incorrectly classified documents. Four tables indicate the distribution of misclassifications across the following parameters: number of sentences, document length, number of discriminating types, and radii. A fifth table outputs overall summary information including Swets' coefficients, and Cleverdon's ratios.

Classification of a set of sample documents upon which the discriminant coefficients are based or a set of independent test documents can be accomplished.

Complete descriptions of these programs including flowcharts and operating instructions are contained in a separate report, Computer Program Description and Instructions, November 22, 1965.

Section IV

EXPERIMENTAL DATA BASE

The data base used in these experiments consisted of 2754 abstracts from the solid state physics field, published by the Cambridge Communications Corporation (CCC), Cambridge, Massachusetts. The abstracts had been classified by CCC into subject categories. Correct experimental classification consisted of matching the CCC classification.

The CCC subject structure contained five categories at the uppermost level and up to ten levels of detail in some categories. Categories selected for the experiment contained at least 280 documents, twice the largest sample size. This yielded three categories at the upper level and three categories at the second level (referred to as the lower level throughout this report) as shown in Figure 2. All documents assigned to these categories constituted the document corpus, consisting of 1743 documents at the upper level and 863 at the lower level. Random sampling was used to draw documents from the document corpus. Three different samples of 35, 70 and 140 documents drawn from each category were used to compute the statistics required by the multiple discriminant technique. The total number of tokens and unique types in each category and group is shown in Table I(a) (upper level category) and Table I(b) (lower level category). The average document length in each category is also shown.

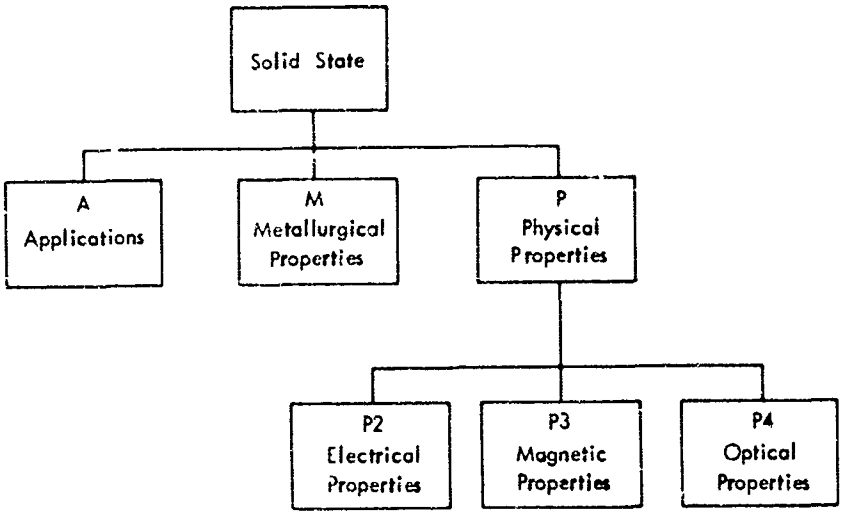


Figure 2. Experimental Solid State Structure

Table I(a) Upper Level Category Statistics

Category	Sample Size	Total Tokens	Total Types	Average Document Length
A	35	2,870	871	82.0
M	35	3,389	938	96.8
P	35	3,858	989	110.2
Group	105	10,117	1,850	96.4
A	70	5,664	1,281	80.9
M	70	6,696	1,383	95.7
P	70	7,198	1,374	102.8
Group	210	19,558	2,549	93.1
A	140	11,583	1,819	82.7
M	140	13,367	1,955	95.5
P	140	14,316	1,860	102.3
Group	420	39,266	3,425	93.5

Table I(b) Lower Level Category Statistics

Category	Sample Size	Total Tokens	Total Types	Average Document Length
P2	35	3,653	917	104.4
P3	35	3,505	873	100.1
P4	35	3,187	844	91.1
Group	105	10,345	1,634	98.5
P2	70	7,273	1,332	103.9
P3	70	7,855	1,366	112.2
P4	70	6,455	1.202	92.2
Group	210	21,583	2,282	102.8
P2	139*	14,523	1,812	103.7
P3	140	15,360	1,815	109.7
P4	140	13,393	1,753	95.7
Group	419	43,276	3,031	103.0

*A processing error caused the omission of one document.

Section V

EXPERIMENTS

Eight experiments were performed to acquire data on the effects of the various parameters. These experiments fall naturally into three groups. The first group consists of experiments on parameters of individual documents and includes:

Experiment 1: Selection of Discriminating Word Types vs Number of Sample Documents

Experiment 2: Category Definition vs Number of Sample Documents

Experiment 3: Classification Effectiveness vs Document Length

Experiment 4: Type-Token Distribution vs Document Length

Experiment 5: Number of Concepts in a Document vs Number of Word Types

The second group consists of experiments on parameters of the classification structure and includes:

Experiment 6: Classification Effectiveness vs Homogeneity of Categories

Experiment 7: Classification Effectiveness vs Level of the Structure

The third group consists of one experiment testing possible measures of classification effectiveness:

Experiment 8: Classification Effectiveness Measures

The following descriptions of each experiment include: purpose, hypotheses, test statistic, experimental procedure, and results. Discussions of the meaning of these results and the conclusions affecting further work are contained in Sections VI and VII.

EXPERIMENT 1: SELECTION OF DISCRIMINATING WORD TYPES vs NUMBER OF SAMPLE DOCUMENTS

The number of sample documents influences classification effectiveness in two ways: first, in the selection of the subset of types for retention as the discriminating set; and second, in the estimation from the sample documents of the population mean vector and dispersion matrix. This experiment reports data on the first effect. Experiment 2 reports on the second effect.

The number of discriminating words selected is limited computationally by a requirement to solve for the eigenvectors of a square matrix of that order. Thus, the need exists to select a subset of types such that the subset will yield the best discrimination between categories, and such that an adequate number of types will occur in each document to be classified. A univariate statistic, the ratio of the pooled among-category variance to the pooled within-category variance, estimates the discrimination ability of each type. The mean frequency of the type within a category estimates the coverage ability of the type. (It is uneconomical to retain a type with high discrimination ability that is relatively rare.) The specific selection criteria are discussed in detail in Section III, "Operational Description."

Hypcthesis

There is no significant difference in the ranking of word types by the discriminant ratio, F as the number of sample documents is increased.

Test Statistic

Spearman's correlation coefficient r_s is used to test for a significant difference in the ranking[9]. The hypothesis will be rejected if the computed value of t is less than the tabulated value of t at $\alpha = .01$. α is the probability level below which the hypothesis will be rejected. The test statistic is

$$r_s = 1 - \frac{6 \sum_{i}^{n} d_i^2}{n(n^2 - 1)} \qquad (15)$$

$$t = r_s \sqrt{\frac{n - 2}{1 - r_s^2}} \qquad (16)$$

where d_i is the difference in the ranks and n is the number of words.

Procedure

To observe the effect of a change in the number of documents, two sets of experiments were performed using sample sizes of 35, 70 and 140 documents — one at the upper level consisting of categories A, M, and P, and one at the lower level consisting of P2, P3, and P4. Random samples of 35, 70, and 140 documents were drawn from each category. The ratio of the pooled among-category variance to the pooled within-category variance was computed for each type. Each type was sorted and ranked according to its ratio. The rank correlation coefficient was computed on the difference in rank of the first 50 types in the 35 document sample and 140 document sample, and between the 70 document sample and the 140 document sample.

Results

Table II lists sets of 50 discriminating words for various sample sizes. The value of t for 50 degrees of freedom at $\alpha = 0.05$ is 2.40. The computed value for t exceeded 2.40. Therefore, for all cases the hypothesis is rejected. Since no significant correlation existed among the sets of discriminating words for various sample sizes, the set based on the 140 document sample was used for the other experiments.

Tables III (a), (b), and (c) contain the discriminant ratio and mean frequencies of the first 100 discriminating words at the upper level for sample sizes 35, 70 and 140. The considerable similarity among the lists of words indicates that this classification parameter warrants further testing. The importance of the selection of discriminating words is discussed further in Section VI, "Homogeneous Categories."

The selection criteria also considers the coverage ability of each type. Since coverage statistics were not available, the mean frequency of each type was employed to indicate the coverage. The validity of using the mean as an indicator of the coverage was tested with a product moment correlation coefficient. A correlation coefficient of 0.9 between the mean frequencies in the 140-document sample and the coverage percentages of test set indicated that the mean was a valid indicator of the coverage. A further test of coverage was made on the percentage of documents in which each type occurred. A χ^2 test revealed that no significant difference existed between the coverage percentages of the sample and test sets. Table IV shows the coverage percentages for the test documents at both the upper and lower levels.

Table II. Discriminating Word Sets for Various Sample Sizes

LOWER LEVEL			UPPER LEVEL		
35	70	140	35	70	140
SPIN	MAGNET	MAGNET	OF	DESCRI	CIRCUI
MAGNET	RESONA	SPIN	EFFECT	CIRCUI	DESCRI
RESONA	SPIN	ABSORP	DESCRI	TRANSI	OPERAT
ABSORP	ABSORP	RESONA	CIRCUI	CRYSTA	CRYSTA
RESIST	PHOSPH	FERROM	ELECTR	DESIGN	FIELD
LIGHT	FERROM	SATURA	DETERM	BAND	OUTPUT
ELECTR	SEMICO	PHOSPH	DISLOC	OF	TRANSI
CONDUC	PARAMA	EMISSI	CRYSTA	UTILIZ	DEVICE
ACTIVA	LIGHT	PARAMA	GROWTH	SIGNAL	K
FERRIM	EMISSI	EXCHAN	OUTPUT	VOLTAG	DESIGN
PHOSPH	CARRIE	FERRIT	REPORT	OUTPUT	UTILIZ
EMISSI	SATURA	SUPERC	DESIGN	SWITCH	CONDUC
PARAMA	RESIST	IRON	CONDUC	USES	OF
EDGE	IRON	INTERA	UTILIZ	CURREN	MAGNET
IRON	EXCHAN	RESIST	ZONE	EMPLOY	PROVID
EXCITA	OPTICA	LIGHT	CHARAC	DIFFUS	SIGNAL
N	BANDS	ANTIFE	ACCURA	FIELD	GROWTH
NEEL	ELECTR	SEMICO	FOUND	REPORT	DISLOC
WAVELE	SUPERC	LUMINE	EXPERI	ELECTR	SWITCH
SATURA	HALL	BAND	TRANSI	USED	VOLTAG
EQUALS	LUMINE	OPTICA	FORMAT	DEVICE	RESONA
LUMINE	SUSCEP	HALL	CONTRO	EFFECT	OSCILL
PEAKS	MOPILE	EXCITA	ONLY	SYSTEM	INPUT
OPTICA	BAND	ELECTR	DIODES	DISLOC	SPIN
FERROM	N	MOBILI	SUPPLY	THEORY	CONTRO
ELEMEN	FERRIT	WAVELE	PROVID	RESONA	ELECTR
O	IS	EDGE	DENSIT	MAGNET	CONNEC
FLUORE	SURFAC	ANISOT	TEMPER	AT	MELT
FERRIT	ACTIVA	CARRIE	RESULT	SURFAC	EMPLOY
SYSTEM	ANTIFE	CONDUC	BETWEE	ABSORP	MEASUR
METHOD	EDGE	CURREN	IN	K	EFFECT
WITHIN	WAVELE	RELAXA	AMPLIF	DIODES	WAS
EVALUA	RELAXA	SUSCEP	S	ON	RESULT
KCL	CENTER	BANDS	BASE	GROWTH	AMPLIF
SENSIT	EXCITA	SPINS	PROCED	USE	ABSORP
HEISEN	G	SURFAC	EDGES	WERE	PULSES
AL	SULFID	RADIAT	GAP	PULSE	FORMAT
ANTIFE	TITANA	MICRON	KINETI	TEMPER	REPORT
DELTA	SPINS	IONS	LATTIC	COUNTE	DIODE
REGION	GREEN	DENSIT	EXPLAI	IN	RECORD
SUSCEP	TRANSV	INTENS	AT	SUPPLY	WERE
POSITI	SILICO	ULTRAV	DEVICE	PARAMA	INFORM
GARNET	CDS	NICKEL	CONSID	AMPLIF	PULSE
DI	INTERA	FLUORE	WAS	CARRIE	DIGITA
APPLIC	HAMILT	FE	FIELD	RAY	IN
EXPONE	MEASUR	SPECTR	IMPURI	CONCEN	SUPPLY
WERE	CONDUC	WIDTHS	IT	BATTER	OBSERV
DENSIT	WIDTH	SULFID	INSTRU	RECORD	FOUND
ANISOT	BARIUM	SPLITT	SAMPLE	EQUALS	TEMPER
OTHER	RECOMB	ISOTRO	USE	BY	INSTRU

Table III(a). Mean Frequencies of Discriminating Words, 35-Document Sample

Word	Category				Word	Category			
	F	A	M	P		F	P2	P3	P4
OF	0.483	3.94	7.17	7.0	ETCHIN	0.110	0.0	0.14	0.0
EFFECT	0 366	0.06	0.14	0.77	PHONON	0.110	0.0	0.0	0.14
DESCRI	0 347	0.69	0 20	0.14	PREVIO	0.110	0.0	0.0	0.14
CIRCUI	0.306	0 77	0.0	0.0	SWITCH	0.110	0.14	0.0	0.0
ELECTR	0.266	0 40	0.34	1.49	SOLID	0.108	0.03	0.17	0.0
DETERM	0.258	0.0	0.46	0.06	DEPEND	0.106	0.06	0.23	0.46
DISLOC	0.252	0.0	0.74	0.03	INTERA	0.104	0.0	0.03	0.23
CRYSTA	0.250	0.09	1.57	0.80	CURREN	0.104	0.46	0.03	0.49
GROWTH	0.249	0.0	0.29	0.0	PERFOR	0.103	0.23	0.0	0.06
OUTPUT	0.228	0.37	0.0	0.03	MECHAN	0.103	0.03	0.14	0.34
REPORT	0.217	0.0	0.34	0.37	BY	0.101	0.74	1.54	1.06
DESIGN	0.206	0.37	0.0	0.03	TECHNI	0.100	0.06	0.31	0.09
CONDUC	0.196	0.09	0.02	0.57	OPERAT	0.100	0.26	0.03	0.09
UTILIZ	0.182	0.20	0.03	0.0	MOLECU	0.100	0.0	0.0	0.20
ZONE	0.172	0.0	0.31	0.0	THAT	0.100	0.43	0.89	1.06
CHARAC	0.164	0.34	0.03	0.11	POSSIB	0.099	0.20	0.09	0 0
ACCURA	0.162	0.29	0.0	0.0	PROCES	0.099	0.03	0.11	0.29
FOUND	0.157	0.0	0.34	0.51	VOLTAG	0.098	0.60	0.03	0.26
EXPERI	0.153	0.09	0.11	0.37	WHICH	0.098	0.83	0.37	0 86
TRANSI	0.151	0.77	0.03	0.31	GROWN	0.098	0.0	0.23	0.0
FORMAT	0.148	0.0	0.26	0.03	PULSE	0.095	0.34	0.0	0.0
CONTRO	0.147	0.37	0.06	0.03	CARRIE	0.035	0.03	0.03	0.26
ONLY	0.145	0.09	0.03	0.29	THERMA	0.092	0.0	0.09	0.29
DIODES	0.143	0.34	0.0	0.0	CONTAI	0.092	0.06	0.23	0.03
SUPPLY	0.143	0.17	0.0	0.0	SIMILA	0.091	0.03	0.09	0.23
PROVID	0.139	0.20	0.03	0.0	HAMILT	0.091	0.0	0.0	0.09
DENSIT	0.137	0.0	0.31	0.03	IDENTI	0.091	0.0	0.0	0.09
TEMPER	0.135	0.14	0.69	0.77	HEATIN	0.091	0.0	0.09	0.0
RESULT	0.133	0.14	0.54	0.69	WHILE	0.091	0.0	0.09	0.0
BETWEE	0.133	0.09	0.09	0.43	PURIFI	0.091	0.0	0.09	0.0
IN	0.128	1.80	2.86	3.17	CAUSED	0.091	0.0	0.09	0.0
AMPLIF	0.128	0.34	0.0	0.03	PORTAB	0.091	0.09	0.0	0.0
S	0.125	0.0	0.0	0.11	ADVANT	0.091	0.09	0.0	0.0
BASE	0.125	0.11	0.0	0.0	ASSUMP	0.091	0.0	0.0	0.09
PROCED	0.125	0.0	0.11	0.0	OCCUR	0.091	0.0	0.0	0.09
EDGES	0.125	0.0	0.11	0.0	FIVE	0.091	0.09	0.0	0.0
GAP	0.125	0.0	0.0	0.11	VARIAB	0.091	0.09	0.0	0.0
KINETI	0.125	0.0	0.11	0.0	RELAXA	0.091	0.0	0.0	0.17
LATTIC	0.120	0.0	0.03	0.23	MEMORY	0.091	0.17	0.0	0.0
EXPIAI	0.120	0.06	0.0	0.23	MELT	0.091	0.0	0.09	0.0
AT	0.120	0.20	0.83	0.97	GROW	0.091	0.0	0.09	0.0
DEVICE	0.117	0.20	0.0	0.0	ALTERN	0.091	0.0	0.0	0.09
CONSID	0.115	0.11	0.06	0.40	TESTS	0.091	0.09	0.0	0.0
WAS	0.115	0.26	1.00	0.66	PULLED	0.091	0.0	0.09	0.0
FIELD	0.115	0.31	0.0	0.63	BRANCH	0.091	0.09	0.0	0.0
IMPURI	0.113	0.0	0.34	0.11	RELIAB	0.091	0.09	0.0	0.0
IT	0 112	0.17	0.40	0.66	FURNAC	0.091	0.0	0.09	0.0
INSTRU	0.111	0.17	0.0	0.0	CAN	0.091	0.40	0.26	0.09
SAMPLE	0.111	0.03	0.17	0.54	SYSTEM	0.090	0.43	0.11	0.09
USE	0.111	0.14	0.03	0.0	DEFORM	0.090	0.0	0.37	0.0

Table III(b). Mean Frequencies of Discriminating Words, 70-Document Sample

Word	Category				Word	Category			
	F	A	M	P		F	P2	P3	P4
DESCRI	0.626	0.84	0.21	0.19	SYMMET	0.091	0.01	0.01	0.14
CIRCUI	0.406	0.81	0.0	0.0	DEPEND	0.091	0.07	0.24	0.43
TRANSI	0.324	0.84	0.07	0.24	ANISOT	0.088	0.0	0.03	0.17
CRYSTA	0.235	0.07	1.26	0.91	GERMAN	0.087	0.0	0.39	0.09
DESIGN	0.193	0.26	0.0	0.01	GROWN	0.085	0.0	0.14	0.0
BAND	0.188	0.01	0.0	0.36	FERROM	0.085	0.0	0.0	0.14
OF	0.183	4.26	6.26	6.77	EXPERI	0.084	0.06	0.17	0.29
UTILIZ	0.182	0.20	0.0	0.01	INTERA	0.084	0.0	0.03	0 16
SIGNAL	0.181	0.44	0.0	0.01	MOBILI	0.084	0.0	0.01	0 24
VOLTAG	0.176	0.46	0.0	0.11	WAVE	0.081	0.06	0.0	0.20
OUTPUT	0.174	0.31	0.0	0.0	ACCOUN	0.081	0.0	0.03	0.11
SWITCH	0.169	0.36	0.0	0.0	CONTRO	0.080	0.26	0.04	0.03
USES	0.164	0.14	0.0	0.0	DIGITA	0.080	0.13	0.0	0.01
CURREN	0.161	0.47	0.0	0.17	M	0.080	0.0	0.0	0.21
EMPLOY	0.158	0.21	0.03	0.0	PULSES	0.079	0.17	0.0	0.04
DIFFUS	0.156	0.0	0.64	0.04	OPERAT	0.079	0.23	0.03	0.09
FIELD	0.153	0.14	0.03	0.64	PROVID	0.079	0.20	0.01	0.03
REPORT	0.141	0.01	0.21	0.30	C	0.076	0.09	0.57	0.30
ELECTR	0.138	0.31	0.44	1.07	PURPOS	0.076	0.07	0.0	0.0
USED	0.135	0.31	0.14	0.03	INSTRU	0.076	0.14	0.0	0.0
DEVICE	0.133	0.21	0.0	0.0	WIDE	0.075	0.07	0.0	0.0
EFFECT	0.124	0.10	0.17	0.70	SPIN	0.075	0.0	0.0	0.23
SYSTEM	0.123	0.51	0.13	0.07	METHOD	0.074	0.09	0.36	0.17
DISLOC	0.122	0.0	0.27	0.03	INFORM	0.073	0.19	0.03	0.01
THEORY	0.122	0.04	0.06	0.31	TAPE	0.073	0.16	0.0	0.0
RESONA	0.120	0.04	0.0	0.27	ROOM	0.073	0.0	0.06	0.17
MAGNET	0.118	0.66	0.01	0.74	ATOMS	0.072	0.0	0.17	0.04
AT	0.117	0.223	0.93	1.06	TREATM	0.072	0.01	0.21	0.04
SURFAC	0.115	0.07	0.51	0.10	SENSIT	0.071	0.20	0.04	0.01
ABSORP	0.113	0.0	0.03	0.36	FOUND	0.071	0.07	0.30	0.31
K	0.110	0.0	0.26	0.57	RELATI	0.069	0.03	0.23	0.09
DIODES	0.110	0.16	0.01	0.0	THAT	0.069	0.34	0.73	0.76
ON	0.108	0.33	0.86	0.79	WINDIN	0.069	0.11	0.0	0.0
GROWTH	0.105	0.0	0.37	0.01	MELTIN	0.069	0.0	0.11	0.0
USE	0.105	0.29	0.07	0.09	PHONON	0.069	0.0	0.0	0.11
WERE	0.103	0.07	0.51	0.44	VARIOU	0.068	0.04	0.03	0.17
PULSE	0.101	0.21	0.0	0.03	DIODE	0.068	0.09	0.0	0.0
TEMPER	0.100	0.17	0.61	0.86	DISSIP	0.068	0.09	0.0	0.0
COUNTE	0.099	0.11	0.0	0.0	READ	0.068	0.09	0.0	0.0
IN	0.099	1.84	2.71	3.11	STRUCT	0.068	0.01	0.24	0.17
SUPPLY	0.096	0.13	0.0	0.0	SPEED	0.067	0.09	0.0	0.01
PARAM	0.096	0.0	0.0	0.13	ADVANT	0.067	0.09	0.0	0.01
AMPLIF	0.096	0.30	0.01	0.0	SOLID	0.067	0.0	0.17	0.02
CARRIE	0.096	0.07	0.10	0.40	STORE	0.067	0.10	0.0	0.0
RAY	0.095	0.01	0.13	0.0	VALENC	0.067	0.0	0.0	0.10
CONCEN	0.095	0.01	0.46	0.31	AMPLIT	0.066	0.11	0.0	0.03
BATTER	0.092	0.09	0.0	0.0	FREQUE	0.066	0.36	0.0	0.30
RECORD	0.091	0.20	0.0	0.01	AL	0.066	0.01	0.21	0.01
EQUALS	0.091	0.01	0.16	0.41	INTERP	0.065	0.0	0.01	0.10
BY	0.091	0.63	1.29	0.86	NICKEL	0.065	0.01	0.0	0.10

Table III(c). Mean Frequencies of Discriminating Words, 140-Document Sample

Word	Category				Word	Category			
	F	A	M	P		F	P2	P3	P4
CIRCUI	0.382	0.86	0.0	0.01	LOGIC	0.084	0.19	0.0	0.0
DESCRI	0.341	0.74	0.27	0.15	ZONE	0.084	0.0	0.18	0.02
OPERAT	0.291	0.46	0.03	0.03	PLANES	0.081	0.0	0.12	0.01
CRYSTA	0.291	0.09	1.46	0.69	INTERA	0.081	0.0	0.02	0.16
FIELD	0.240	0.12	0.04	0.75	AT	0.079	0.35	0.92	1.06
OUTPUT	0.212	0.45	0.0	0.01	SYSTEM	0.078	0.41	0.19	0.06
TRANSI	0.192	0.69	0.05	0.27	IMPURI	0.078	0.0	0.28	0.09
DEVICE	0.182	0.28	0.01	0.01	GROWN	0.077	0.0	0.20	0.01
K	0.177	0.04	0.05	0.79	REQUIR	0.076	0.21	0.04	0.04
DESIGN	0.176	0.32	0.01	0.01	EXPERI	0.076	0.11	0.20	0.39
UTILIZ	0.173	0.17	0.01	0.0	USE	0.075	0.25	0.08	0.05
CONDUC	0.167	0.06	0.04	0.46	INTERF	0.075	0.02	0.15	0.0
OF	0.162	4.16	6.28	6.43	DEVELO	0.075	0.14	0.03	0.02
MAGNET	0.162	0.48	0.02	0.84	COMPOS	0.074	0.01	0.21	0.04
PROVID	0.157	0.34	0.03	0.03	DENDRI	0.073	0.0	0.16	0.0
SIGNAL	0.156	0.25	0.0	0.02	POWER	0.071	0.34	0.02	0.14
GROWTH	0.150	0.01	0.49	0.04	DEPEND	0.071	0.09	0.16	0.39
DISLOC	0.147	0.0	0.59	0.04	CLOSE	0.071	0.0	0.08	0.0
SWITCH	0.136	0.37	0.0	0.02	CURREN	0.069	0.44	0.03	0.14
VOLTAG	0.134	0.55	0.01	0.14	DIODES	0.068	0.13	0.0	0.0
RESONA	0.133	0.06	0.0	0.33	MELTIN	0.068	0.0	0.13	0.0
OSCILL	0.123	0.26	0.0	0.01	COUNTE	0.068	0.25	0.01	0.01
INPUT	0.120	0.18	0.01	0.0	RELIAB	0.068	0.09	0.01	0.0
SPIN	0.114	0.0	0.0	0.34	BETWEE	0.068	0.18	0.16	0.41
CONTRO	0.109	0.31	0.09	0.01	SPEED	0.067	0.11	0.0	0.03
ELECTR	0.109	0.28	0.46	0.99	COMPUT	0.067	0.11	0.0	0.03
CONNEC	0.108	0.17	0.01	0.01	SUITAB	0.066	0.09	0.01	0.0
MELT	0.105	0.0	0.15	0.0	CALCUL	0.065	0.01	0.13	0.20
EMPLOY	0.104	0.19	0.03	0.03	THEORY	0.065	0.03	0.09	0.22
MEASUR	0.104	0.28	0.37	0.79	PERFOR	0.065	0.14	0.01	0.04
EFFECT	0.103	0.09	0.25	0.53	APPLIE	0.063	0.06	0.04	0.21
WAS	0.103	0.19	0.86	0.61	RECEIV	0.063	0.11	0.01	0.0
RESULT	0.102	0.12	0.40	0.47	READIN	0.063	0.07	0.0	0.0
AMPLIF	0.102	0.30	0.01	0.04	PULLED	0.063	0.0	0.07	0.0
ABSORP	0.099	0.0	0.03	0.19	ATTRIB	0.063	0.0	0.0	0.07
PULSES	0.098	0.12	0.0	0.0	CAPABL	0.063	0.07	0.01	0.0
FORMAT	0.095	0.0	0.15	0.02	USED	0.061	0.31	0.11	0.14
REPORT	0.095	0.07	0.27	0.33	IMPEDA	0.061	0.08	0.0	0.0
DIODE	0.094	0.16	0.0	0.0	PORTAB	0.060	0.06	0.0	0.0
RECORD	0.092	0.16	0.0	0.01	BECOME	0.060	0.0	0.0	0.06
WERE	0.091	0.11	0.60	0.33	GROWIN	0.060	0.0	0.09	0.0
INFORM	0.090	0.14	0.01	0.0	RELAXA	0.060	0.01	0.01	0.19
PULSE	0.088	0.39	0.01	0.02	DISCUS	0.060	0.29	0.57	0.46
DIGITA	0.088	0.12	0.0	0.01	PARAMA	0.060	0.01	0.0	0.08
IN	0.087	1.79	2.62	3.03	ASSUME	0.060	0.0	0.02	0.09
SUPPLY	0.087	0.15	0.0	0.0	FIELDS	0.059	0.04	0.0	0.16
OBSERV	0.086	0.09	0.29	0.51	CONVER	0.059	0.13	0.01	0.01
FOUND	0.085	0.06	0.36	0.35	BASE	0.059	0.09	0.01	0.0
TEMPER	0.084	0.23	0.57	0.87	DIFFUS	0.058	0.04	0.39	0.14
INSTRU	0.084	0.14	0.0	0.0	HOLE	0.057	0.01	0.0	0.11

Table IV. Percentage of Documents Containing Specific Type

Upper Level		Lower Level	
Type	% of Documents Containing Type	Type	% of Documents Containing Type
CIRCUI	0.099	SUPERC	0.061
DESCRI	0.298	RESIST	0.117
OPERAT	0.089	SEMICO	0.106
OUTPUT	0.054	HALL	0.057
TRANSI	0.190	ELECTR	0.439
DEVICE	0.046	MOBILI	0.064
DESIGN	0.059	CARRIE	0.131
PROVID	0.062	CONDUC	0.157
SIGNAL	0.050	CURREN	0.095
SWITCH	0.042	SURFAC	0.111
VOLTAG	0.087	DENSIT	0.077
OSCILL	0.041	RECOMB	0.062
CONTRO	0.064	TYPE	0.148
AMPLIF	0.038	FIELD	0.374
PULSE	0.048	SCATTE	0.061
SYSTEM	0.125	METHOD	0.146
CRYSTA	0.337	MAGNET	0.347
GROWTH	0.059	SPIN	0.157
DISLOC	0.044	RESONA	0.160
SURFAC	0.130	FERROM	0.092
METHOD	0.161	SATURA	0.081
FOUND	0.197	PARAMA	0.084
IMPURI	0.085	EXCHAN	0.050
DISCUS	0.378	FERRIT	0.056
DIFFUS	0.064	IRON	0.058
STRUCT	0.096	INTERA	0.123
SINGLE	0.181	ANTIFE	0.046
CONCEN	0.109	ANISOT	0.083
C	0.153	RELAXA	0.088
CONTAI	0.077	SUSCEP	0.056
OXYGEN	0.037	IONS	0.085
TECHNI	0.091	MOMENT	0.062
FIELD	0.206	ABSORP	0.145
K	0.160	PHOSPH	0.053
CONDUC	0.138	EMISSI	0.057
EXPERI	0.207	LIGHT	0.079
MAGNET	0.226	LUMINE	0.037
RESONA	0.088	BAND	0.108
SPIN	0.082	OPTICA	0.083
ELECTR	0.367	EXCITA	0.061
MEASUR	0.356	WAVELE	0.046
EFFECT	0.249	EDGE	0.033
RESULT	0.283	BANDS	0.047
REPORT	0.213	RADIAT	0.050
DEPEND	0.195	MICRON	0.057
OBSERV	0.217	INTENS	0.084
TEMPER	0.353	FLUORE	0.020
BETWEE	0.220	SPECTR	0.146

EXPERIMENT 2: CATEGORY DEFINITION vs NUMBER OF SAMPLE DOCUMENTS

The number of sample documents used to define a category affects the classification performance in many ways. Some of these can be measured directly, while others can be measured only indirectly. In any statistical technique it is naturally assumed that increasing the sample size will yield a better estimate of the parameters. One of the problems of multivariate techniques is that the sample size necessary increases with the number of variables. Since the population statistics are not known, the parameters derived from the largest sample were considered to be estimates of the true statistics. A test for a significant difference between the parameters of the largest sample and smaller samples was performed. When the number of sample documents is large enough, there should be no significant difference in the parameters. Hence misclassification due to poor estimates of the mean and dispersion should be reduced to a minimum.

The purpose of this experiment is not to show how much misclassification will be caused by an error in the estimate of the parameters, but to determine whether an error exists in the estimate of the parameters defining the category. When the number of sample documents is increased to the point where there is no significant difference in the estimate of the parameters, it can be assumed that a larger number of sample documents would not affect the estimated values and therefore, that no further gain in classification performance could be achieved.

Hypotheses:

H_1: There is no significant difference in the mean vector of a category based on two different sample sizes.

H_2: There is no significant difference in the dispersion matrix of a category based on two different sample sizes.

Test Statistic

For H_1 the test statistic is Kullback's $2\hat{I}^{(10)}$. The value of the statistic is compared to a χ^2 table. The hypothesis will be rejected if $2\hat{I} > \chi^2$ at $\alpha = 0.01$.

For H_2 the test statistic is also Kullback's $2\hat{I}$. The value of the statistic is compared to Fisher's table of β^2. The hypothesis will be rejected if $2\hat{I}$ is greater than the table value at $\alpha = 0.01$.

Procedure

The mean vector and dispersion matrix of each category were computed and transformed to the reduced space. The statistics of two categories at a time were compared.

For the test of H_1, consider the categories independent samples with n_1 and n_2 independent observations (documents) from k-variate (words) normal populations. The test statistic is

$$2\hat{I} = (\bar{X}_1 - \bar{X}_2)^T \left(\frac{1}{n_1} S_1 + \frac{1}{n_2} S_2\right)^{-1} (\bar{X}_1 - \bar{X}_2). \qquad (17)$$

and

$$\text{d.f.} = (r-1)\,K,$$

where

r is the number of samples.

For the test H_2, consider the categories as independent samples with n_1 and n_2 observations (documents) from k-variate (words) normal populations. The test statistic is

$$2\hat{I} = N_1 \log \frac{|S|}{|S_1|} + N_2 \log \frac{|S|}{|S_2|} \qquad (18)$$

where

$$S = \frac{N_1}{N} S_1 + \frac{N_2}{N} S_2$$

$$N_1 = n_1 - 1$$

$$N_2 = n_2 - 1$$

and

$$N = N_1 + N_2.$$

Enter the β^2 table with values of β^2 and the degrees of freedom computed by the following expressions:

$$\beta^2 = \frac{(2k^3 + 3k^2 - k)}{12} \left(\frac{1}{N_1} + \frac{1}{N_2} - \frac{1}{N}\right) \tag{19}$$

$$\text{d.f.} = \frac{k(k+1)}{2} .$$

Results

The experiment was conducted on all the categories at both the upper and lower levels. The sample sizes were 70 and 140 documents in all cases.

The value of X^2 at $\alpha = 0.01$ is 9.2. Table V shows that a stable estimate of the mean vectors was obtained for some of the categories. The experiment tested for a significant difference between samples of 70 and 140, which means when the null hypothesis is accepted that a sample of 70 is sufficient. The test offered no information concerning the 140-document sample itself. Information concerning the 140-document sample could be obtained by testing it against a larger sample or all the available documents.

Table V. Test of Category Means

	Upper Level			Lower Level		
	A	M	P	P_2	P_3	P_4
$2\hat{I}$ decision	1.11 accept	24.71 reject	1.67 accept	1.68 accept	39.68 reject	10.53 reject

The value of β^2 for acceptance of the hypothesis at $\alpha = 0.01$ is 11.3. Table VI shows that considerable difference existed between the dispersions. The dispersion matrices in the original variable space contain 48 x 48 elements, and thus may require a larger number of documents to obtain a stable estimate.

It is interesting to note that, as the number of sample documents increased, the dispersion increased. While this might imply that classification performance would decrease as the number of sample documents increased, additional information on this parameter reported in Section VI indicates that improvement in classification performance can be expected as sample size is increased.

Table VI. Test of Category Dispersions

	Upper Level			Lower Level		
	A	M	P	P_2	P_3	P_4
$2\hat{I}$ decision	286.4 reject	295.3 reject	271.5 reject	316.5 reject	310.2 reject	345.8 reject

EXPERIMENT 3: CLASSIFICATION EFFECTIVENESS vs DOCUMENT LENGTH

The purpose of this experiment was to assess the effect of variation of document length on classification performance and to give some indication of the minimum length that the technique can handle. Since the data base was confined to abstracts with a range of 20 to 250 tokens, considerable data was collected concerning very short documents and the variation in this range.

Hypotheses

H_1: There is no significant difference in the mean number of tokens in each document between the correctly classified documents and the incorrectly classified documents in each category.

H_2: There is no significant difference in the mean number of sentences in each document between the correctly classified documents and the incorrectly classified documents in each category.

Test Statistic

Student's t was used to test the hypotheses for equality of the two sample means, with equal variances, and unequal sample sizes. The test statistic is:

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{s_{\bar{x}_1 - \bar{x}_2}} \qquad (20)$$

$$s_{\bar{x}_1 - \bar{x}_2} = \sqrt{s^2\left(\frac{n_1 + n_2}{n_j\, n_2}\right)} \qquad (21)$$

$$s^2 = \frac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2}{(n_1 - 1) + (n_2 - 1)} \qquad (22)$$

The hypotheses will be rejected if the computed value of t is greater than the tabulated value of t at $\alpha = 0.01$.

Procedure

The mean number of tokens and the standard error of the mean of the correctly and incorrectly classified documents in each category was computed. Let x_1 and x_2, s_1^2 and s_2^2, and n_1 and n_2 denote the sample means, variances, and sizes, respectively. The sets of correctly and incorrectly classified documents were considered to be independent samples from normal populations having a common but unknown variance. An F test was used to establish that the variances were equal.

The degrees of freedom $(n - 1)$ for each category varied from 124 to 864. Since the change in t in this range was very slight, the value of t for 100 degrees of freedom was used for all categories. The value of t for 100 degrees of freedom at $\alpha = 0.01$ is 2.58.

The experiment was conducted on the sample and test documents of all the categories at both the upper and lower levels. The number of sample documents used to compute the discriminant coefficients was 140.

Results

A summary of the results concerning H_1 is shown in Table VII. The results indicated that there was no significant difference in the means. Differences as large as those obtained can be attributed to sampling variation alone. Therefore, in this set of documents, the length of the document does not have a very significant effect on performance. The influence of document length on other parameters that do significantly effect performance is discussed in Section VI.

A summary of the results concerning H_2 is shown in Table VIII. The value of t for this test is also 2.58. The results indicated that there was no significant difference in the means. Therefore, in this set of documents, the number of sentences appeared to have no significant effect on performance. The insignificant difference between the means of the sample set and the test set indicates that the samples were very representative of their populations.

Table VII. Test of Document Length

	Upper Level			Lower Level		
	A	M	P	P_2	P_3	P_4
Sample						
Correct, $\bar{x}_1$	83.2	96.0	105.8	103.6	111.0	97.1
Incorrect, $\bar{x}_2$	79.9	93.2	82.0	101.4	89.3	88.5
t	0.40	0.34	2.2	0.14	1.3	0.95
Decision	accept	accept	accept	accept	accept	accept
Test						
Correct, $\bar{x}_1$	79.5	95.6	108.4	101.6	108.1	95.6
Incorrect, $\bar{x}_2$	85.3	80.3	92.7	108.7	82.1	88.3
t	0.93	2.4	4.5	0.72	1.8	0.76
Decision	accept	accept	reject	accept	accept	accept

Table VIII. Test of Number of Sentences

	Upper Level			Lower Level		
	A	M	P	P_2	P_3	P_4
Sample						
Correct, $\bar{x}_1$	4.5	5.3	5.5	5.1	5.5	5.2
Incorrect, $\bar{x}_2$	4.3	4.9	4.3	4.7	4.9	4.5
t	0.48	0.89	2.3	0.66	0.68	1.4
Decision	accept	accept	accept	accept	accept	accept
Test						
Correct, $\bar{x}_1$	4.5	5.1	5.5	5.3	5.5	5.1
Incorrect, $\bar{x}_2$	4.8	4.3	4.9	5.4	4.3	4.5
t	0.89	2.3	3.3	0.30	2.1	1.1
Decision	accept	accept	accept	accept	accept	accept

EXPERIMENT 4: TYPE-TOKEN DISTRIBUTION vs DOCUMENT LENGTH

One problem encountered in many textual data handling applications is determining how many documents are needed to get the system going. The problem has two aspects: cost and statistical reliability. Expressions exist for computing the sample size required for a given confidence level when the distribution of a random variable is known. However, when using this technique the distribution of words is not known and other techniques must be employed.

In addition to the lack of a distribution function, textual applications frequently have the missing variable problem: the same words do not occur in every document. This experiment was concerned with the interrelationship of the number of documents and their lengths and the occurrence of word types.

One of the steps of the classification technique is the selection of a subset of words on which classification is based. The selection criteria is based on a univariate statistic computed from a sample of documents. The sample must be large enough to ensure not only that the statistic is reliable but also that all the significant words have occurred. If the sample is too small, words that are truly good discriminators may not have appeared yet.

The appearance of new word types may be analyzed by their arrival rate. The arrival rate may be measured either on a per document basis or on a per token basis. For the statistical technique to be feasible, the arrival rate must decrease as the number of documents or the number of tokens increases. In fact, for the technique to be economical the arrival rate must approach a small number fairly rapidly. The arrival rate will never approach zero, since many documents will contain a type that is used only in that document. This is not a problem, however, as long as a subset of words can be found, each of which occurs in some portion of the documents, and that the subset as a whole offers sufficient coverage of each document.

If one has long documents it would appear that not as many documents are required to obtain an adequate sample size as with short documents. But, the relative terms "long" and "short" need some quantitative bounds. The range of documents in the solid state data base is from 20 to 250 tokens. The purpose of this experiment was to determine if any significant difference exists in the number of types obtained from samples of long and short documents exists. The documents were grouped into two sets: those having 100 or more tokens and those having less than 100 tokens.

This experiment provided information on the interrelation of the length of documents and the number of documents, the cumulative distribution of unique types, and the arrival rate of new types.

Hypothesis

There is no significant difference in the relationship of the cumulative number of types to the cumulative number of tokens when the individual document lengths are changed.

Test Statistic

χ^2 is used to test for a significant difference between the cumulative types for short documents and the cumulative types for long documents.

Procedure

Figure 3 shows the subset of the structure selected to perform this experiment. Categories containing a sufficiently large number of documents were chosen. To minimize variation caused by different subject matter, subcategories of major categories were selected. All the documents of each category were then partitioned into two sets. Set A contained documents of 100 word tokens or more. Set B contained documents of less than 100 word tokens.

The words in each document were counted and the cumulative number of tokens and types was computed. Linear interpolation was performed on the data where necessary in order to make the comparisons. χ^2 was computed on the differences between values of the short and long documents. The degree of freedom is one less than the number of documents being compared.

Results

The data in Table IX indicates that no significant difference existed between long and short documents at the third and fourth levels and that a significant difference existed between long and short documents at the first and second levels. The difference existing at the upper levels may be due to variation of two parameters not controlled by this experiment. At the first level documents were drawn from five different categories, while at the P levels documents were drawn only from the P category. It could be assumed that a relation exists between the number of different categories and the number of different words. The other parameter is the variation in average document length. The average length of documents in A, M, P is 83, 96, and 102 tokens, respectively. This deviation would cause more A documents to be placed in the below-100 set and more P documents to be placed in the above-100 set.

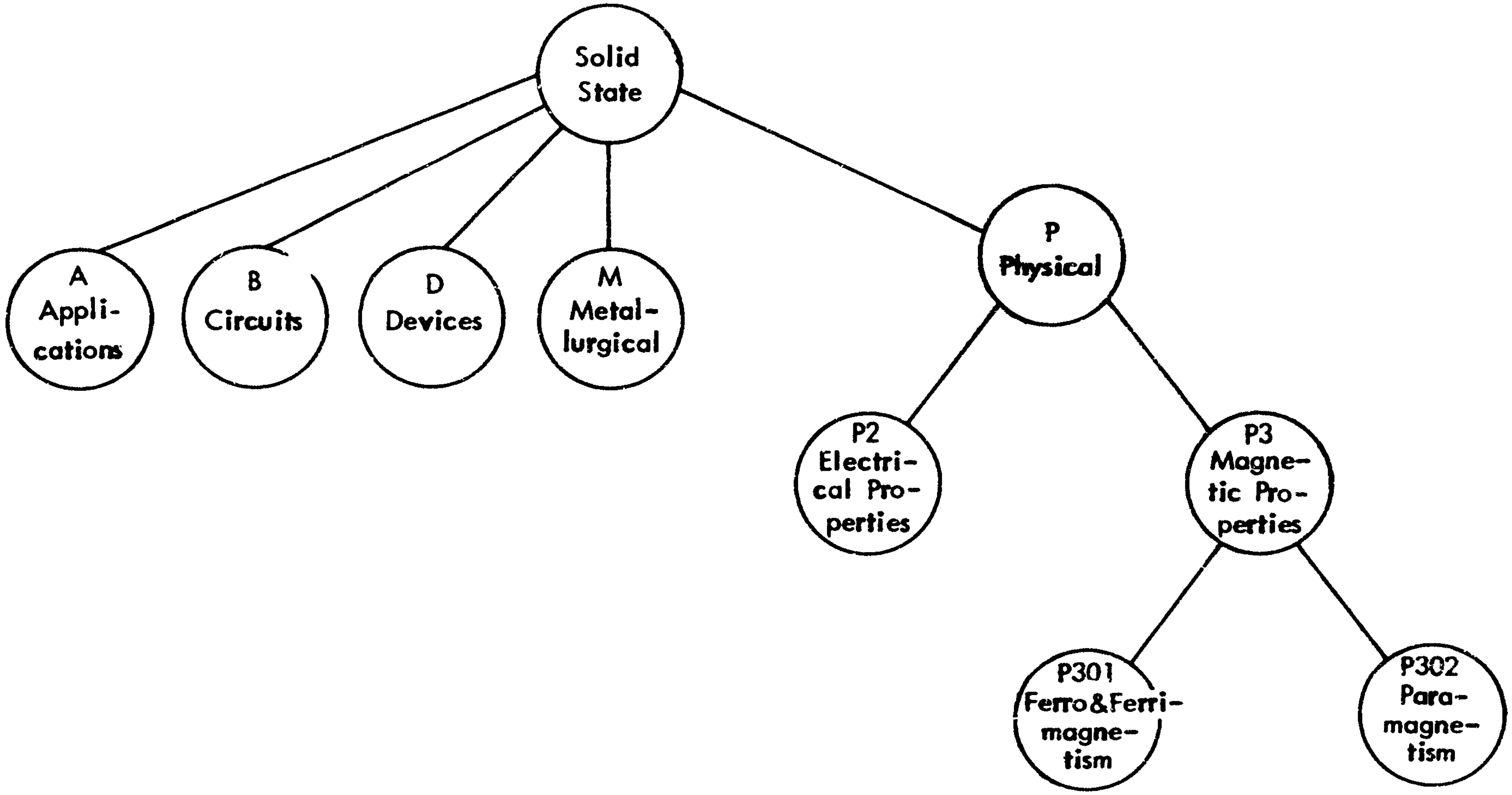


Figure 3. Experimental Structure for Type-Token Test

Table IX. Cumulative Word Types

Total Tokens	Category											
	P301A	P301B	P302A	P302B	P2A	P2B	P3A	P3B	PA	PB	SSA	SSB
500	225	232	228	239	214	234	224	251	232	236	243	257
600	257	267	265	271	258	273	274	285	263	274	281	293
700	294	301	299	299	287	304	298	312	294	303	311	326
800	326	336	332	322	323	333	327	346	320	332	340	354
900	355	363	355	342	357	350	353	375	347	359	372	384
1,000	385	395	378	373	390	380	379	401	377	385	398	419
1,500	509	531	494	489	519	507	500	502	517	519	525	572
2,000	623	626	591	585	624	622	607	632	654	639	637	694
2,500	734	754	679	672	714	710	691	714	737	729	745	789
3,000	810	825	742		798	782	808	794	821	812	852	862
3,500	846	891	803		881	863	866	866	902	898	947	962
4,000	910	960	887		931	918	918	925	954	994	1009	1039
4,500	971	1004	949		1015	976	969	983	1012	1053	1085	1112
5,000	1041		1003		1072	1050	1029	1043	1067	1134	1167	1173
5,500	1102		1069		1110	1110	1082	1083	1119	1174	1219	1248
6,000	1164		1123		1151	1160	1150	1135	1169	1238	1274	1339
6,500	1206		1197		1198	1217	1187	1193	1208	1300	1331	1425
7,000	1255		1296		1259	1272	1240	1227	1257	1353	1369	1492
7,500	1299		1338		1301	1326	1287	1275	1343	1396	1412	1565
8,000	1349		1377		1343	1353	1335	1332	1382	1454	1448	1628
8.500	1390				1372	1402	1373		1429	1496	1493	1671
9,000	1438				1403	1439	1403		1460	1546	1540	1714
9,500	1475				1450	1466	1435		1490	1579	1572	1755
10,000	1504				1495	1500	1482		1523	1618	1615	1800
15,000					1836		1810		1861	1927	1969	2173
20,000					2102				2119	2159	2305	2507
25,000									2343	2386	2541	2791
30,000									2525	2599	2828	3018
35,000									2677		3010	3217
40,000									2851		3192	3377
45,000									2997		3363	3554
50,000									3142		3495	3699
55,000									3262		3626	3829
60,000									3356		3752	3967
65,000									3498		3890	4084
70,000											4057	4214
75,000											4146	4359
80,000											4253	4469
85,000											4336	4603
90,000											4445	4724
95,000											4541	4850
100,000											4638	

d.f.	12	8	23	19	27	40
Computed X^2	9.75	1.91	10.58	12.262	56.05	402.18
Table X^2	21.03	15.51	35.17	30.14	40.11	55.80
Decision	Accept	Accept	Accept	Accept	Reject	Reject

Additional data was obtained which substantiates the conclusion of no significant difference for homogeneous categories. The cumulative distributions of P3A and PA were compared with the distribution from a random sample of 140 documents in each category. The random sample was drawn independent of document length. The chi-square values were less than the criterion, and the hypothesis of no significance was therefore accepted.

The experiment also revealed that the shape of the cumulative distribution curve is fairly independent of the subject category. The only variation was regular, viz., in moving up the structure toward a more general category the number of types increased slightly. The cumulative number of types is therefore a function of the total number of tokens observed rather than the number of documents.

From the SSB column of Table IX it can be observed that the arrival rate of new types decreased considerably. In the first 10,000 tokens 1800 types had arrived, whereas in the 80,000 to 90,000 interval only 255 types had arrived. Since the average SSB document length is 72, and each document contains 1.83 new types, 96 percent of the types in each document has already been observed. The arrival rate is approaching a small number, thus confirming that the words selected from a sample will be representative and will not adversely affect classification performance.

EXPERIMENT 5: NUMBER OF CONCEPTS IN A DOCUMENT vs NUMBER OF WORD TYPES

The number of concepts in a docu ent is important in a classification system, since documents discussing more than one subject can be placed only in the category dealing with the major or primary subject, or in several categories, one for each subject. If documents dealing with more than one subject have a greater number of word types for a given document length, then the probability of classifying them correctly into several categories based on a fixed set of discriminating words will be greater, since more of the discriminating words can be expected to appear. On the other hand, if the number of word types for a given document length is the same for single-and multiple-concept documents, then longer multiple-concept documents will be needed to achieve the same classification accuracy as with single-concept documents.

This experiment, therefore, was designed to investigate the relationship of the number of word types in the document to the number of concepts in the document. The number of concepts in the document was taken as the number of categories assigned to it by Cambridge Communications Corporation.

Hypothesis

For a given number of tokens, there is no significant difference in the number of types between singly-classified documents and multiply-classified documents.

Test Statistic

Since the type-token distribution is not known, the non-parametric U Test was used(11). This test is based on the sums of ranks and is useful in testing for the equality of two populations. Let $X_1, X_2, \ldots, X_m$ be a sample from a population with frequency function f(X), and let $Y_1, Y_2, \ldots, Y_n$ be a sample from a population with frequency function f(Y). Then, to test the hypothesis that f(X) = f(Y), rank the combined set of values X_i ($i=1, 2, \ldots, m$) and Y_j ($j=1, 2, \ldots, n$), in order of increasing magnitude. Let T denote the sum of the ranks of the Y values. Then the statistic

$$U = mn + \frac{n(n+1)}{2} - T \tag{23}$$

is approximately normally distributed when m and n are both greater than 8. Furthermore,

$$E\ [U] = \frac{mn}{2} \tag{24}$$

$$\sigma_U^2 = \frac{mn\ (m+n+1)}{12}\ . \tag{25}$$

Procedure

Thirteen cases were tested to establish the hypothesis. In each case, a fixed document length was chosen, and documents of that length were examined to see whether they were singly, doubly or triply classified. Comparisons were planned to be made when there were eight or more documents in all sets. However, because of the difficulty of finding enough documents of the same length in the relatively small data base being used, fewer than eight documents were accepted in some cases.

After the documents had been selected for each case, they were ranked, and the values of U. E [U] and σ_U^2 were computed. The hypothesis was accepted at the 1 percent level as long as:

$$X = \left| \frac{U - E\ [U]}{\sigma_U} \right|^2 < (2.58)^2 = 6.65 \qquad (26)$$

Results

The data used and results obtained are displayed in Table X. In all cases tested, the hypothesis was verified: for documents of the same length, there was no difference in the number of types between documents with different numbers of concepts.

The mean document length and mean number of types for singly, doubly, and triply classified documents were computed and are given below:

Class	Mean Length	Mean No. of Types	Types/Length
Singly classified	93.13	61.07	0.66
Doubly classified	96.35	62.18	0.65
Triply classified	110.43	69.15	0.63

Thus, although the density of types is not significantly different among the three classes of documents, the lengths do increase with the number of concepts, and the number of types increases as would be expected in longer documents. Also, as expected, the ratio of the two drops as the length increases. However, the wide range in document lengths used produces a large variance, so the increase in mean lengths is not significant.

EXPERIMENT 6: CLASSIFICATION EFFECTIVENESS vs HOMOGENEITY OF CATEGORIES

Lack of homogeneity of a category represented by the statistical model may be caused by several diverse factors. The factors involve the definition of the subject category, the documents assigned to it, the words used by the author, the words selected as the discriminating subset, and the estimates of the statistical parameters defining that category.

The conceptual definition of a category, such as Solid State Devices, may include concepts also included in the definition of another category. The documents assigned to a category may also be assigned to another category and thus contain words representative of more than one category. Therefore the mean

Table X. Number of Word Types vs Number of Concepts

Case	1			2			3			4			5			6			7			8			9			10			11			12			13		
Number of Tokens	51			71			92			92			92			94			104			105			111			121			121			121			129		
Number of Types in Single-Concept Documents	38 39 42 39 38 38	40 43		49 51 52 56 58 41	47 53 54 18 56 40	47 50 50 52	56 61 62 62 66 67			56 61 62 62 66 67						64 66 72 53 60 53	61 62 65 68 66		59 62 68 62 66 68	75 61 75 59 59 69	72 74	70 74 63 61 67 58	73		81 87 61 62 70 66	78 65 71 78		74 78 79 63 65 79	83 74 80 64 73 77		74 78 79 63 65 79	83 74 80 64 73 77					67 82 77 78 79 80	81 82 84 86 93	
Number of Types in Double-Concept Documents	38 12 29 34 40 39	39 39		50 52 53 54 54 54	60 50 51		65 68 57 69 57 59	68					57 57 59 65 68 68	60		56 63 69 69 71 60	62 63 61 62 66 67								65 63 81 70 75			78 79 82 73 78 67	70					78 79 82 73 78 67	70		76 60 72 72 74 75	82	
Number of Types in Triple-Concept Documents										62 60 60 61 63 66			62 60 60 61 63 66						69 69 56 65 67			72 57 58 62 70 74									78 72 83 68 71			78 72 83 68 71					
λ	0 47			2 25			0 32			0 23			0. 08			0. 24			0. 10			0. 13			0. 03			0 04			0. 44			0 06			5 02		
Accept Hypothesis	Yes			Yes			Yes			Yes			Yes			Yes			Yes			Yes			Yes			Yes			Yes			Yes			Yes		

frequencies among categories may not differ significantly in some cases. The effect of these factors in the model should be reflected in the estimates of the mean, dispersion, and position of the ellipses representing the categories.

The purpose of this experiment was to determine if any decrease in classification performance could be attributed to lack of homogeneity as measured by the magnitude of the dispersion matrix.

Hypothesis

There is no significant difference in the ranking of the classification effectiveness and the homogeneity of a category.

Test Statistic

Spearman's correlation coefficient r_s is used to test for a significant difference in ranking.

The test statistic is given by equations (15) and (16). The hypothesis will be rejected if the computed value of t is less than the tabulated value of t at $\alpha = 0.01$.

Procedure

The determinant of the dispersion matrix of each category in the discriminant space was computed. The categories were ranked in descending order of their determinants. The percentage of correct classifications among the test documents for each category was computed, and the categories were ranked in order of increasing accuracy. For this test, categories from the upper and lower levels were considered together. The classification results were based on discriminant coefficients computed from the 140-document sample.

Results

Table XI. Test of Homogeneity

Rank_d	Category	Determinant	Effectiveness	Rank_e
1	A	0.1011	0.88	5
2	P	0.0994	0.73	1
3	M	0.0494	0.73	2
4	P3	0.0333	0.92	6
5	P2	0.0332	0.87	4
6	P4	0.0299	0.85	3

The value of t with 4 degrees of freedom at $\alpha = 0.01$ is 4.6. The hypothesis was rejected since the computed value of t is 0.17. Thus, the relative size of the dispersion matrix in itself appeared to have no significant affect on performance.

The determinants of all the lower level categories were considerably smaller than those of the upper level. Two factors that are independent of level contributing to the dispersion matrix were higher frequencies and words occurring in more than one category. An important factor related to homogeneity causing a decrease in performance appeared to be the position of each word in the discriminant space. Some of the words were extremely far from the centroids of their categories and in fact very close to the centroid of another. Influence of the position of words on performance is discussed in Section VI, "Homogeneity of Categories."

EXPERIMENT 7: CLASSIFICATION EFFECTIVENESS vs LEVEL OF THE STRUCTURE

This experiment was intended to test the ability of the technique to classify at more than one level. Since the model is stated in general terms, it should be insensitive to a change in level. The technique should be useful at several levels of detail, requiring only that a sufficient number of sample documents exist to estimate the statistical parameters.

Hypothesis

There is no significant difference in the probable error matrix from level to level.

Test Statistic

χ^2 is used to test for a significant difference between 3 x 3 tables of observed frequencies and expected frequencies. The number of degrees of freedom is 6.

Procedure

The probable error matrix was computed for all categories at both levels. The sample size was 140 documents for each category. The χ^2 test requires integers, therefore the percentages were multiplied by 100.

To compare observed frequencies with expected frequencies, some similarity must exist between the populations. No such similarity existed between subject headings at one level and those at another level. Therefore, the statistical properties of the categories were tested. Since the sequence of categories is arbitrary (P2, P3, P4; P4, P3, P2), the lower level categories were assigned the following sequence for this test: P3, P2, P4. The criteria of ordering was based on the location of the categories in the discriminant space. The first category is the one which is on an axis by itself; the second category is that one of the remaining two nearest the origin.

Results

Table XII. Test of Lower Level

Upper Level				Lower Level			
	A	M	P		P3	P2	P4
A	1.00	0.00	0.00	P3	1.0	0.00	0.00
M	0.01	0.93	0.06	P2	0.0	0.91	0.09
P	0.03	0.03	0.93	P4	0.0	0.01	0.99

The value of χ^2 at $\alpha = 0.05$ for six degrees of freedom is 11.07. The hypothesis is accepted since the computed χ^2 of 7.260 is less than 11.07.

This test did not provide adequate information about the effects of level of the structure on effectiveness. Therefore, in addition to this test, the lower level categories were included in many other experiments. Corresponding results between levels are shown in the tables in the Appendix. In general, better results were achieved at the lower level than at the higher level because of the set of discriminating words used. The position of the discriminating words in the discriminant space appears to have a more significant effect on classification performance than the level.

EXPERIMENT 8: CLASSIFICATION EFFECTIVENESS MEASURES

Measures of classification effectiveness are needed in the following phases: in the testing phase not only to measure overall classification effectiveness but also to measure changes in effectiveness due to a change in a parameter; in the implementation phase to test whether a new set of data meets the assumptions required for the technique; and in the operational phase to continually monitor the system performance.

This experiment tested two measures of classification effectiveness. One is the probable error matrix which requires the assumption that the data is multivariate normal, and the other is the actual error matrix of the sample documents which does not require this assumption. The value of the selected measure would be to indicate the expected error in classifying new documents and to aid in making implementation decisions concerning a specific set of data.

Hypotheses

H_1: There is no significant difference between the probable error matrix and the test document error matrix.

H_2: There is no significant difference between the sample document error matrix and the test document error matrix.

Test Statistic

χ^2 is used to test for a significant difference between 3 x 3 tables of observed frequencies and expected frequencies. The number of degrees of freedom is six.

Procedure

The probable error matrix was computed by considering the centroid representing a category as a random point in the discriminant space. The probability of membership in each category was computed by equation (14). In the ideal case, the probability of its being in its own category would be 1.0. Thus all the diagonal elements of the matrix would be 1.0 and all the off-diagonal elements would be 0.0. In the real case, the off-diagonal elements would indicate the amount of overlap of categories and hence the probable error of misclassifying into that category.

The sample and test error matrices were derived from the actual number of documents classified by the technique. Since the number of documents in each category in the sample set and test set differ, they have been normalized by the number of the documents in the category. The χ^2 test requires integers, therefore the percentages were multiplied by 100.

Results

Table XIII. Error Matrices

	Upper Level								
	Probable Error			Sample Error			Test Error		
	A	M	P	A	M	P	A	M	P
A	1.00	0.00	0.00	0.88	0.03	0.09	0.88	0.05	0.07
M	0.01	0.93	0.06	0.04	0.82	0.14	0.06	0.73	0.21
P	0.03	0.04	0.93	0.05	0.10	0.85	0.08	0.18	0.73

	Lower Level								
	Probable Error			Sample Error			Test Error		
	P2	P3	P4	P2	P3	P4	P2	P3	P4
P2	0.91	0.0	0.09	0.94	0.04	0.03	0.87	0.05	0.08
P3	0.00	1.0	0.00	0.04	0.94	0.01	0.06	0.92	0.02
P4	0.01	0.0	0.99	0.16	0.01	0.83	0.14	0.01	0.85

Table XIV. χ^2 Values

	Upper Level	Lower Level
Probable Error	27.90 reject	39.61 reject
Sample Error	12.13 accept	5.43 accept

The value of χ^2 at $\alpha = 0.05$ for six degrees of freedom is 11.07. The results shown in Table XIV reject the hypothesis that the probable error matrix is a good predictor of performance, and accept the hypothesis that the sample error matrix is a good predictor of performance. The results were very encouraging because they indicated that results obtained from the sample set were representative of those to be expected from the test set. In previous work, results obtained from the sample set were usually much higher than those from the test set.

Although the probable error matrix cannot be used as a predictor for the sample or test set, it could have considerable value. The probable error matrix indicated the amount of overlap caused by the specific set of discriminating words used. The discrepancies in the off-diagonal elements between the probable error and sample error matrices were due to the dispersion of the documents in the discriminant space. They were more widely dispersed than the estimate of the dispersion matrix indicated.

Section VI

ANALYSIS OF CLASSIFICATION EFFECTIVENESS

The effectiveness of the classification technique and the causes of misclassifications are analyzed here and related to the significant classification parameters. The analysis of the effect of variation in the level of the structure is discussed in connection with each parameter. Some new parameters which appear to be adversely affecting classification performance are identified. Classification performance is reported in terms of several efficiency and effectiveness measures. Detailed data on the classification parameters and performance are recorded in the tables in the Appendix and are discussed in Section V.

NUMBER OF SAMPLE DOCUMENTS (EXPERIMENTS 1 AND 2)

Classification of an independent test set of documents, as well as the sample documents themselves, was performed on random samples of 35, 70 and 140 documents in each category at both the upper and lower levels. Two sets of discriminating words were used, one at each level. At each level the set of discriminating words remained fixed, while the number of sample documents was changed. Each discriminating word set contained 48 words; 16 words represented each category. The total population of documents at the upper level was 1753 and at the lower level was 872. The number of test documents for a particular run was the total population minus those drawn to form the sample. Table XV shows the percentage of documents classified into each category. For example, 88 per cent of the test documents manually preclassified into category A were classified automatically into category A based on a 140-document sample. Five per cent of the A documents were misclassified into M and seven per cent were misclassified into P.

As the sample size increases, the percentage of correct test documents increases while the percentage of correct sample documents decreases. However, the percentages appear to be approaching a common point within each category as shown in Figure 4. In category A of the upper level the test percentages are: 0.79, 0.84 and 0.88, while the sample percentages are: 0.97,

Table XV. Classification Performance

Upper Level

	Sample Documents								
Sample Size	35			70			140		
Effectiveness Actual/Auto	A	M	P	A	M	P	A	M	P
A	0.97	0.03	0.0	0.91	0.04	0.04	0.88	0.03	0.09
M	0.03	0.89	0.09	0.03	0.81	0.16	0.04	0.82	0.14
P	0.0	0.06	0.94	0.06	0.16	0.78	0.05	0.10	0.85
	Test Documents								
Sample Size	35			70			140		
Effectiveness Actual/Auto	A	M	P	A	M	P	A	M	P
A	0.79	0.05	0.17	0.84	0.10	0.07	0.88	0.05	0.07
M	0.13	0.64	0.23	0.09	0.74	0.17	0.06	0.73	0.21
P	0.11	0.21	0.68	0.08	0.30	0.62	0.08	0.18	0.73

Lower Level

	Sample Documents								
Sample Size	35			70			140		
Effectiveness Actual/Auto	P2	P3	P4	P2	P3	P4	P2	P3	P4
P2	0.94	0.0	0.06	0.96	0.01	0.03	0.94	0.04	0.03
P3	0.0	1.0	0.0	0.04	0.94	0.01	0.04	0.94	0.01
P4	0.06	0.0	0.94	0.07	0.01	0.91	0.16	0.01	0.83
	Test Documents								
Sample Size	35			70			140		
Effectiveness Actual/Auto	P2	P3	P4	P2	P3	P4	P2	P3	P4
P2	0.71	0.10	0.19	0.80	0.07	0.13	0.87	0.05	0.08
P3	0.13	0.80	0.07	0.05	0.89	0.06	0.06	0.92	0.02
P4	0.26	0.07	0.67	0.16	0.01	0.83	0.13	0.01	0.85

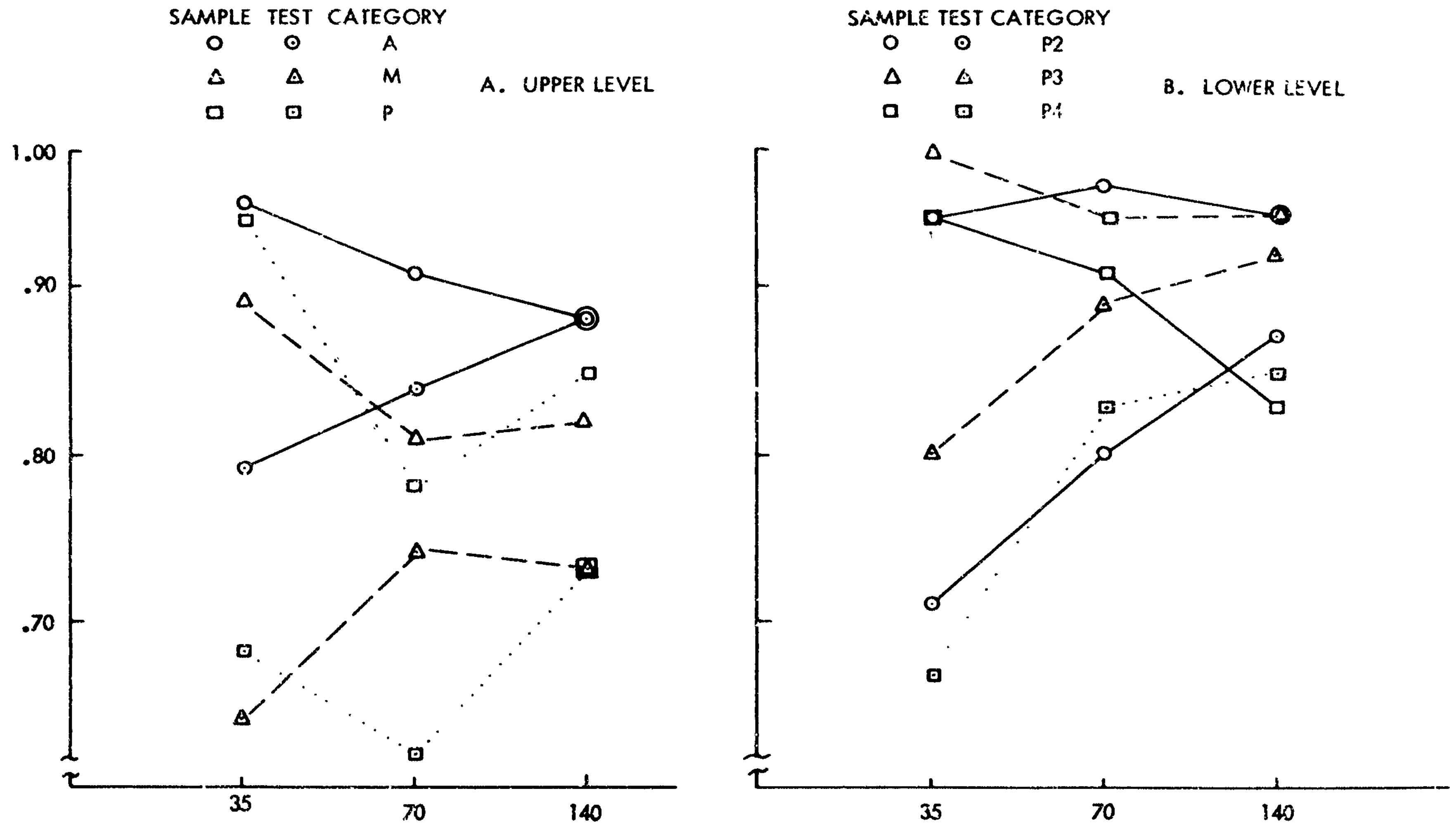


Figure 4. Effect of Classification Performance Due to the Number of Sample Documents

0.91 and 0.88. In category P3 of the lower level the test percentages are: 0.80, 0.89 and 0.92 while the sample percentages are: 1.0, 0.94 and 0.94. The dispersion of the category's mean increases as sample size increases. When it is small, the sample documents classify well but are not representative of the whole population; hence the test documents do not classify as well. As the dispersion increases, the sample set is more representative of the population and, even though the sample documents perform less satisfactorily, they appear to be representing the population more faithfully. An increase in sample size beyond 140 would not appear to offer any significant increase in classification performance.

Experiment 2 attempted to test the effect on classification performance of an increase in the number of reference documents. To test the hypothesis, one statistic was selected, the determinant of the category's dispersion. It was assumed that more documents would be misclassified in categories having larger dispersion. Although Table XV indicates that better results were obtained with an increase in the number of sample documents, the hypothesis of Experiment 2 must be rejected on the basis of the category's dispersion. Apparently the magnitude of a category's dispersion is not a good indicator of the percentage of misclassifications to be expected.

The number of sample documents also affects classification performance through the selection of the subset of discriminating word types, and Experiment 1 was conducted to study this. Selected words have a high discriminating coefficient and a high likelihood of occurring in many documents. If words are selected from too small a sample, they may not have a high discriminating coefficient nor a high likelihood of occurring in the population. Therefore, these statistics were computed on samples of 35, 70 and 140 documents. Experiment 1 indicated that, according to the test used, samples of 35 and 70 were too small to assure reliable estimates of the discriminating and likelihood statistics. However, the great similarity between the lists, the fact that the rank of a word on the list of selected words is not important, and the general effectiveness of the technique even when based on a 35- or 70-document sample, are indications that the present method of word selection is adequate until further experimentation on this phase of the technique can be conducted.

DOCUMENT LENGTH (EXPERIMENTS 3, 4, AND 5)

The effect of document length on classification performance was analyzed in terms of the total number of words (tokens), the number of sentences, and the number of concepts contained in each document.

The purpose of Experiment 3 was to determine if document length adversely affects classification performance. The results of the experiment indicated that document length did not have this effect since no significant difference was found between the mean lengths of correctly and incorrectly classified documents. In data bases involving a greater variation in document length, a significant difference may develop. However, in data bases such as the one being tested, where the variation in document length is relatively small, it did not appear to be a significant parameter. Table 2 in the Appendix contains information on classification performance relative to document length. A slight trend in favor of longer documents exists in categories M and P. However, there are many categories in which short documents classified as well as long documents. No minimum length was found below which documents were consistently misclassified. Experiment 3 also revealed that no significant difference existed in the mean number of sentences between the correctly and incorrectly classified documents.

Variations in other parameters related to document length may have more effect than the document length itself. These include the number of sentences and the number of different types. In addition, the number of discriminating types occurring in a document is dependent on the number of types in a document. Therefore, an additional hypothesis concerning the mean number of discriminating types in correctly and incorrectly classified documents was tested. A significant difference existed between the mean number of discriminating types (7.6 and 5.9) occurring in the correctly and incorrectly classified documents at the upper level, whereas the difference between the means (4.9 and 4.3) at the lower level was found to be insignificant. In fact, 85 per cent of the lower level documents having a mean of 4.9 types were correctly classified, whereas only 76 per cent of the upper level documents having a mean of 7.6 types were correctly classified. A definite correlation between performance and the number of discriminating types can be observed for each category from Table 3 in the Appendix. Classification performance would be increased by increasing the number of types occurring in each document. This can easily be accomplished by increasing the number of discriminating words. The performance level in the current experiments was achieved with only 48 discriminating types.

Experiment 4 showed that the cumulative type-token distribution is not dependent on document length. Hence, one can infer that the total number of words occurring in documents defining a category is more important than the number of documents. Similarly, Experiment 5 indicated that the number of concepts in a document does not affect the number of types occurring; therefore, one probably needs longer multiple-concept documents to achieve the same level of effectiveness obtained with single-concept documents.

HOMOGENEOUS CATEGORIES (EXPERIMENT 6)

Lack of homogeneity in a category can be due to the definition of the subject categories, the sample documents and the discriminating words selected to represent the categories. Experiment 6 revealed that the dispersion of a category is not an adequate measure of its homogeneity. The dispersion of an individual category neglects the amount of overlap between categories and the causes of this overlap. A relation between the overlap of categories and the discriminating words selected to represent it was found to have a direct effect on classification performance. The location of each word in the discriminant space directly influences the location of the category.

Figure 5 shows a plot of the centers for categories A, M, P based on a 140-document sample. The "centours"(12) for each category are the locus of points within which 95 per cent of the documents belonging to that category are expected to lie. Some overlap exists which indicates an ambiguous region. The coordinates of the center of each ellipse are functions of the mean frequencies and the coefficients of the discriminating types. The coordinates of the categories' centroids and dispersion are listed in Table 6 of the Appendix.

Category A has less of its area overlapped by other categories, and achieved 15 per cent more correctly classified documents than either of the other two. Axis I provides the primary discrimination between category A and categories M or P. Axis II provides the primary discrimination between M and P. The plot of the centers indicates the change required to increase classification performance. Categories M and P need to be moved farther apart along axis II and need to be farther from category A along axis I.

The relative positions of the ellipses are functions of the sample documents and the discriminating words. The effect of each word on the position of the ellipse can be observed from Figure 6. The coordinates of each word in Figure 6 are obtained by multiplying its discriminant coefficients by a frequency of one. The 16 words selected to represent each category should be clustered near their category mean. The words SPIN (0.10, - 0.24) and CONDUC (0.07, - 0.38) are examples of good discriminators for category P. The word REPORT (0.17, 0.15) is an example of a poor discriminator for category P. The word REPORT contributes a positive increment to the centroid of category P causing it to overlap category M. The word TRANSI (-0.10, -0.13) causes the category A ellipse to overlap category P. Words having a high coefficient which initially appear to be good discriminators also contribute to misclassification. DESCRI (-0.27, 0.02) and DISCUS (-0.37, 0.26) are examples of this type. These words occurred in documents belonging to all three categories

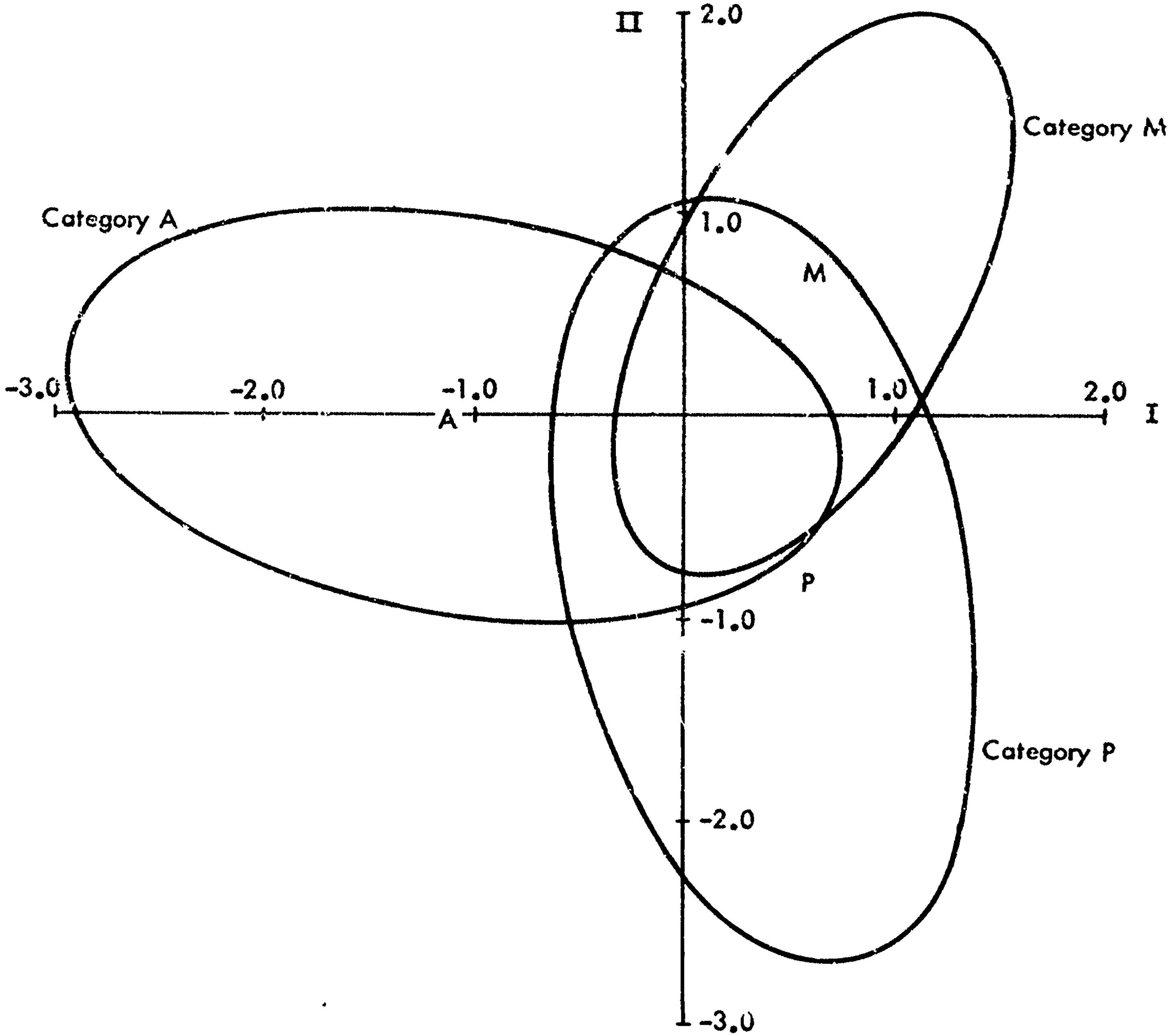


Figure 5. "Centours" of Upper Level Categories, 140-Document Sample

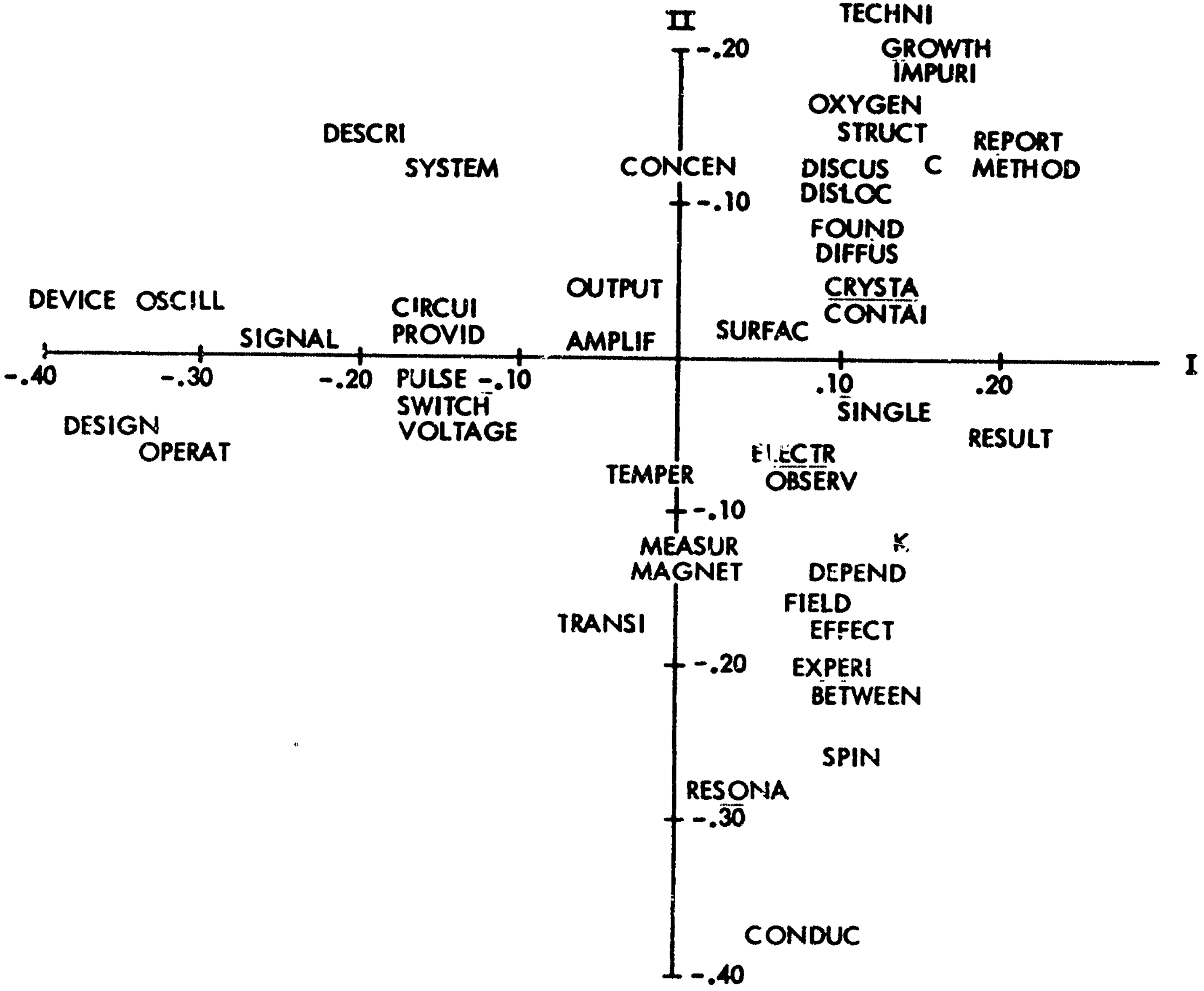


Figure 6. Location of Upper Level Words, 140-Document Sample

and frequently occurred in misclassified documents containing three types or less. The classification performance of documents containing five word types would be increased by the deletion of this type of word from the discriminating word set.

Improvement in classification performance due to the deletion of poor discriminators is indicated by the superior performance of the lower level categories. The discriminating word set for the lower level contained only five candidates for deletion, whereas the upper level contained 14. The discriminant coefficients based on 140-document samples are shown in Table XVI.

Many misclassified documents had small coordinate values which placed them near the origin in the ambiguous region. Improving the homogeneity of categories should reduce the amount of overlap, thereby minimizing the ambiguous region and increasing classification performance.

CLASSIFICATION EFFECTIVENESS MEASURES (EXPERIMENT 8)

It would be desirable to use the probable error matrix as a predictor of classification effectiveness, since then it would be unnecessary to classify all the sample documents. Experiment 8 revealed, however, that the sample error matrix is a better effectiveness measure. This result is no doubt due to the probable error matrix indicating the error to be expected at only one point, the category mean, and not taking the dispersion into account, whereas the sample error matrix does since it is based on many documents.

In addition to these measures of classification effectiveness based on system parameters, it is desirable to consider overall performance measures which allow comparison of this technique to others. Three such measures were therefore computed for each category at both the upper and lower levels. The measures computed were the percentage of correctly classified documents, Swets' ratios and Cleverdon's ratios(13).

Swets' and Cleverdon's ratios are based on the familiar two-by-two contingency table. (See Table XVII.) Swets uses the ratios to compute an additional performance measure, E, which is the statistical operating characteristic capable of showing the amount of type I and II errors. Values of E can be interpreted objectively since it has been widely used in other statistical applications. Curves of the operating statistic are available for the complete range of relevancy from 0.0 to 1.0 (Reference 13). The operating characteristic, in addition to providing a measure of effectiveness, also provides a

Table XVI(a). Normalized Discriminant Coefficients for Upper Level Categories, 140-Document Sample

	I	II		I	II
CIRCUI	-0. 17	0. 02	DIFFUS	0. 09	0. 12
DESCRI	-0. 20	0. 19	STRUCT	0. 09	0. 17
OPERAT	-0. 32	-0. 05	SINGLE	0. 11	-0. 01
OUTPUT	-0. 03	0. 07	CONCEN	-0. 03	0. 13
TRANSI	-0. 10	-0. 13	C	0. 09	0. 14
DEVICE	-0. 41	0. 10	CONTAI	0. 08	0. 07
DESIGN	-0. 36	-0. 05	OXYGEN	0. 06	0. 18
PROVID	-0. 15	0. 01	TECHNI	0. 07	0. 26
SIGNAL	-0. 25	0. 01	FIELD	0. 01	-0. 16
SWITCH	-0. 19	-0. 02	K	0. 03	-0. 14
VOLTAG	-0. 19	-0. 06	CONDUC	0. 07	-0. 38
OSCILL	-0. 32	0. 09	EXPERI	0. 08	-0. 17
CONTRO	0. 04	0. 10	MAGNET	-0. 05	-0. 12
AMPLIF	-0. 04	-0. 02	RESONA	0. 00	-0. 29
PULSE	-0. 14	0. 00	SPIN	0. 10	-0. 24
SYSTEM	-0. 12	0. 17	ELECTR	0. 02	-0. 06
CRYSTA	0. 09	0. 09	MEASUR	-0. 02	-0. 12
GROWTH	0. 13	0. 22	EFFECT	0. 04	-0. 17
DISLOC	0. 05	0. 13	RESULT	0. 18	-0. 06
SURFAC	0. 03	0. 03	REPORT	0. 17	0. 15
METHOD	0. 14	0. 14	DEPEND	0. 04	-0. 15
FOUND	0. 08	0. 12	OBSERV	0. 05	-0. 06
IMPURI	0. 13	0. 21	TEMPER	-0. 03	-0. 07
DISCUS	0. 05	0. 13	BETWEE	0. 10	-0. 20

Table XVI(b). Normalized Discriminant Coefficients for Lower Level Categories, 140-Document Sample

	I	II		I	II
SUPERC	0. 14	-0. 39	IRON	-0. 33	0. 15
RESIST	0. 01	-0. 07	INTERA	-0. 10	0. 03
SEMICO	0. 14	-0. 23	ANTIFE	0. 02	0. 01
HALL	0. 13	-0. 19	ANISOT	-0. 07	0. 07
ELECTR	0. 05	-0. 06	RELAXA	-0. 17	0. 08
MOBILI	-0. 01	-0. 07	SUSCEP	-0. 21	0. 14
CARRIE	0. 06	-0. 04	IONS	-0. 06	-0. 02
CONDUC	0. 03	-0. 14	MOMENT	0. 08	-0. 02
CURREN	0. 01	-0. 05	ABSORP	0. 12	0. 17
SURFAC	0. 05	-0. 10	PHOSPH	0. 12	0. 26
DENSIT	-0. 04	-0. 20	EMISSI	0. 18	0. 27
RECOMB	0. 06	-0. 20	LIGHT	0. 02	0. 17
TYPE	0. 10	-0. 06	LUMINE	0. 11	0. 17
FIELD	0. 00	-0. 03	BAND	0. 01	0. 04
SCATTE	-0. 05	-0. 10	OPTICA	0. 09	0. 16
METHOD	-0. 03	-0. 13	EXCITA	0. 08	0. 11
MAGNET	-0. 23	0. 03	WAVELE	0. 26	0. 37
SPIN	-0. 15	0. 01	EDGE	0. 01	0. 01
RESONA	-0. 22	-0. 04	BANDS	0. 02	-0. 01
FERROM	-0. 29	0. 15	RADIAT	0. 15	0. 19
SATURA	-0. 21	0. 00	MICRON	0. 09	0. 18
PARAMA	-0. 28	-0. 01	INTENS	-0. 03	0. 14
EXCHAN	-0. 29	0. 07	FLUORE	0. 10	0. 07
FERRIT	-0. 32	0. 09	SPECTR	-0. 02	0. 10

measure of the breadth of a particular query. In an ideal system E would be constant, and the breadth of a query could be determined by the slope of the curve at the point resulting from the relevance values of that query.

Table XVII. Pertinency and Retrieval 2 x 2 Contingency Table

	Pertinent	Not Pertinent	Total
Retrieved	a	b	a + b
Not Retrieved	c	d	c + d
Total	a + c	b + d	a + b + c + d

Swets calls his ratios the conditional probability of a hit, $Pr_p(R) = a/a+c$; and the conditional probability of a false drop, $Pr_{\bar{p}}(R) = b/b+d$. The conditional probability of a hit becomes the ordinate of the operating characteristic and the conditional probability of a false drop becomes its abscissa. Figure 7 shows a family of normal operating characteristic curves plotted on a double probability scale. A value of E = 0 indicates a random system regardless of the percentage of hits. A hit ratio of 0.70 and a false drop ratio of 0.70 would yield an E of zero. The greatest value of the operating characteristic occurs in the analysis of misclassification errors and the design of systems. It indicates the direction the analysis should proceed to improve system performance. The values of a system on the curve will indicate whether a greater increase in performance can be expected by increasing the hit ratio or decreasing the false drop ratio. Points of diminishing return can also be found from the curves. Cleverdon calls his ratios recall, a/a+c; relevance, a/a+b; and precision, a+b/a+c. These ratios have limited value in analyzing effectiveness since their theoretical distributions are not known and the size of the document collection is ignored. The element d of the two-by-two table does not enter into any of the computations. Values of Cleverdon's ratios are reported in Table 5 of the Appendix for comparisons with other retrieval techniques.

Classification effectiveness values of the upper and lower level categories are plotted in Figure 7(a). The clustering of the lower level categories indicates a fairly consistent system performance. The effect of a change in sample size is shown in Figure 7(b). Increasing the sample size raised the performance level by increasing the hit ratio and decreasing the false drop ratio. Since all plots are to the left of a diagonal perpendicular to E = 0 line, the point of diminishing returns has not been reached.

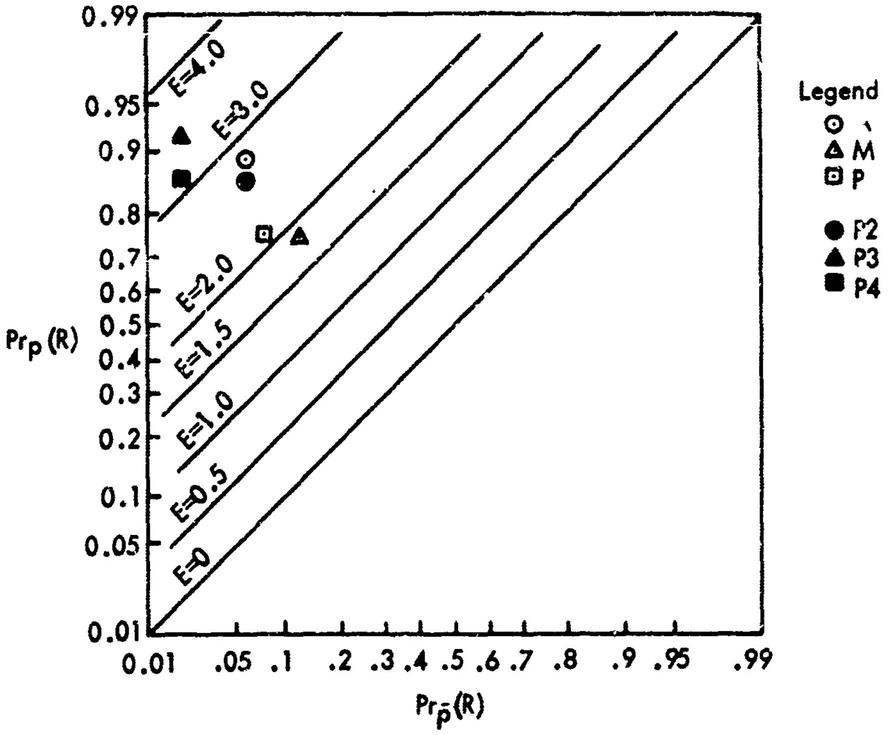


Figure 7(a). Classification Effectiveness vs. Level, Measured by Swets' E

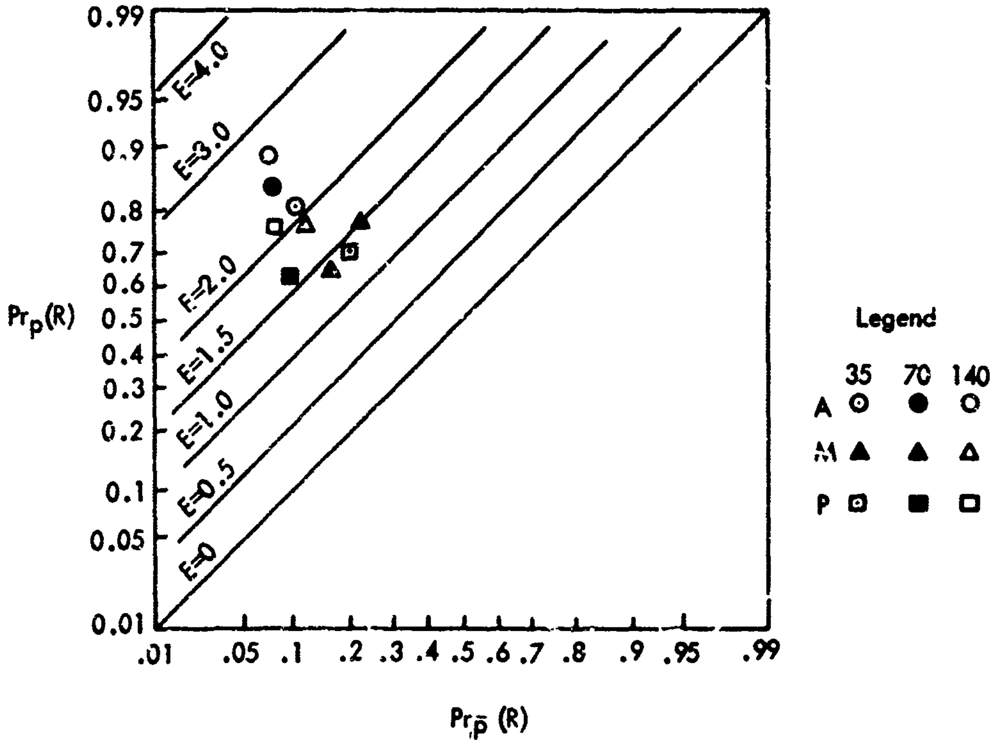


Figure 7(b). Classification Effectiveness vs. Number of Documents, Measured by Swets' E

Section VII

RECOMMENDATIONS

Experiments on the principal classification parameters have produced valuable information concerning their effect on classification performance. Also, new parameters have been identified which appear to have an important relationship to performance. This section discusses the most significant of these and recommends specific areas for further study and experimentation.

Since the specific word types retained in the discriminating word set directly affect the location and size of the category ellipses, experiments should be performed involving the deletion of certain word types from the set based on classification performance. The word selection criteria should be investigated and the possible inclusion of the multivariate discriminant coefficient in the criteria should be examined.

The decision rule employed in the current experiments was based on the χ^2 value of each document. This did not consider the a priori probability distribution of the documents. For example, it was known that there were approximately four times as many P documents as A and M documents in the test population. Use of a probability measure weighted according to the a priori probability of each category may offer a significant improvement. An expression for a probability measure is given by equation (14).

Since there will always be a region of ambiguity in the area of overlapping categories near the origin, documents lying within this region should not be classified. A criterion such as minimum number of word types or minimum radius should be investigated to identify such documents. The use of such a criterion would also yield information concerning the efficiency of the technique. The performance level of the technique on the set of documents meeting the criterion may be sufficiently high to warrant operational testing. If 90 percent of a document corpus could be classified automatically with 95 percent effectiveness, and the remaining 10 percent could be identified automatically, then it would be a minor problem to classify these manually.

A sufficient number of the significant classification parameters have been identified and studied to warrant testing the classification technique on documents selected from an operational environment. Such testing should be directed toward identifying any remaining problems in the application of the technique, as well as toward the further improvement of classificaticn efficiency and effectiveness.

Section VIII

SUMMARY

Classification experiments were pe:formed on a d^ta base of approximately 2700 abstracts from the solid state physics fiel'. The purpose of these experiments was to assess the effect of certain classification parameters on the effectiveness and utility of a document classification technique based upon multiple discriminant functions.

Multiple discriminant functions derived from word frequency information in manually classified documents are used to test whether or not an incoming document should be assigned to a particular category. Since a test is made for each category, documents could be assigned to more than one category, and any number of categories can be handled by the technique. In applications requiring subcategories or a hierarchical structure, the tecnnique would classify documents into the lowest level by making sequential decisions at each level.

Of all the parameters investigated, the number of sample documents used to define a category yielded the greatest increase in classification performance. Experiments were conducted with three different sample sizes: 35, 70, and 140 documents in each category for a total of 105, 210 and 420 documents.

The percentage of correctly classified test documents increased from 0.69 to 0.76 at the upper level and from 0.74 to 0.88 at the lower level. An interesting effect of the opposite nature with respect to the sample documents was observed. As the number of sample documents was increased, the percentage of correctly classified sample documents decreased from 0.93 to 0.85 at the upper level and from 0.96 to 0.90 at the lower level. Although better results were obtained in the sample set using a smaller number of documents, the estimate of the coefficients was apparently not representative enough of the whole population. As the number of sample documents was increased, the percentage of correctly classified sample and test documents approached a common point, indicating that the sample is most representative of the population.

The number of sample documents also affects classification performance through the subset of word types selected to represent the category. Since each word does not occur in every document, its occurrence is dependent on the number of documents sampled. Selection decisions based on samples that are too small may omit words that are truly good discriminators. List of words were ranked according to their univariate discriminant coefficients based on various sample sizes, and a comparison of the first 50 words on each list were made. A rank correlation test revealed that the subset of words from the 35- and 70-document samples were significantly different from the 140-document sample. Therefore, when selecting subsets of words, descriptors or keywords, one must be assured that the sample size is large enough to adequately represent the sample.

The variation in document length did not adversely affect classification performance. The range of the length of the abstracts was approximately 20 to 250 tokens. Tests for a significant difference between the mean document lengths and the mean numbers of sentences in the correctly and incorrectly classified documents revealed no significant differences.

A more significant effect on classification performance was due to the number of discriminating word types occurring in a document. Although documents containing 12 or more of the 48 possible types were usually classified correctly, no minimum below which documents were usually misclassified was found. Misclassification was more often due to the specific discriminating words found in the document, rather than merely the number. The contribution of specific word types to a category's mean and dispersion, and consequent adverse effect on classification performance, was identified and warrants further investigation.

The arrival rate of new word types was found to be approaching a reasonably small number after several hundred documents had been counted. This assured that a subset of discriminating words can be selected on the basis of an economical number of tokens represented by a small number of sample documents.

Considerable variation in performance was found among categories. The primary variation was found to be due to the words selected to represent the categories. Those categories containing words representing more than one category did not perform as well as other categories. An increase in performance in those categories could be achieved by better selection of the words representing them, thereby increasing overall performance.

The ability of the technique to effectively classify at more than one level was demonstrated by the results achieved in several experiments.

Investigation of effectiveness measures revealed that Swets' E appeared to be the best statistic to measure system performance and to analyze the causes of misclassifications.

Experimentation has indicated that classification effectiveness as high as 92 percent can be expected. Classification parameters have been identified which will aid in the analysis of causes of misclassified documents, and provide guidance in the application of this classification technique to new data bases. Sufficiently high performance levels have been achieved to warrant experimental application of the multiple discriminant technique to operational data bases in order to identify any remaining problems, and to further improve classification efficiency and effectiveness.

Appendix

DETAILED DATA ON CLASSIFICATION PARAMETERS AND PERFORMANCE

TABLE 1 EFFECTIVENESS VS NUMBER OF SENTENCES
35 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 2 48 DISCRIMINATING WORDS

SAMPLE DOCUMENTS			CATEGORY A		
NUMBER OF SENTENCES	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
1	0.	0.	0.	0.	0.
2	4.	4.	0.	1.00	0.
3	9.	8.	1.	0.89	0.11
4	7.	7.	0.	1.00	0.
5	3.	3.	0.	1.00	0.
6	6.	6.	0.	1.00	0.
7	3.	3.	0.	1.00	0.
8	2.	2.	0.	1.00	0.
9	1.	1.	0.	1.00	0.
10	0.	0.	0.	0.	0.
11	0.	0.	0.	0.	0.
12	0.	0.	0.	0.	0.
13	0.	0.	0.	0.	0.
14	0.	0.	0.	0.	0.
15	0.	0.	0.	0.	0.
TOTAL	35.	34.	1.	0.97	0.03

	MEAN	S.D.
CORRECT DOCUMENTS	4.62	1.89
INCORRECT DOCUMENTS	3.00	0.

TABLE 1 EFFECTIVENESS VS NUMBER OF SENTENCES
35 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 2 48 DISCRIMINATING WORDS

SAMPLE DOCUMENTS		CATEGORY M			
NUMBER OF SENTENCES	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
1	1.	1.	0.	1.00	0.
2	0.	0.	0.	0.	0.
3	5.	5.	0.	1.00	0.
4	5.	3.	2.	0.60	0.40
5	6.	5.	1.	0.83	0.17
6	6.	6.	0.	1.00	0.
7	8.	8.	0.	1.00	0.
8	3.	3.	0.	1.00	0.
9	0.	0.	0.	0.	0.
10	1.	0.	1.	0.	1.00
11	0.	0.	0.	0.	0.
12	0.	0.	0.	0.	0.
13	0.	0.	0.	0.	0.
14	0.	0.	0.	0.	0.
15	0.	0.	0.	0.	0.
TOTAL	35.	31.	4.	0.89	0.11

	MEAN	S.D.
CORRECT DOCUMENTS	5.45	1.76
INCORRECT DOCUMENTS	5.75	2.41

TABLE 1 EFFECTIVENESS VS NUMBER OF SENTENCES
35 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 2 48 DISCRIMINATING WORDS

SAMPLE DOCUMENTS			CATEGORY P		
NUMBER OF SENTENCES	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
1	1.	0.	1.	0.	1.00
2	2.	2.	0.	1.00	0.
3	5.	5.	0.	1.00	0.
4	4.	4.	C.	1.00	0.
5	5.	4.	1.	0.80	0.20
6	7.	7.	0.	1.00	0.
7	4.	4.	0.	1.00	0.
8	1.	1.	0.	1.00	0.
9	1.	1.	0.	1.00	0.
10	3.	3.	0.	1.00	0.
11	1.	1.	0.	1.00	0.
12	1.	1.	0.	1.00	0.
13	0.	0.	0.	0.	0.
14	0.	0.	0.	0.	0.
15	0.	0.	C.	0.	0.
TOTAL	35.	33.	2.	0.94	0.06

	MEAN	S.D.
CORRECT DOCUMENTS	5.91	2.60
INCORRECT DOCUMENTS	3.00	2.00

TABLE 1 EFFECTIVENESS VS NUMBER OF SENTENCES
35 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 2 48 DISCRIMINATING WORDS

SAMPLE DOCUMENTS		CATEGORY TOTAL			
NUMBER OF SENTENCES	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
1	2.	1.	1.	0.50	0.50
2	6.	6.	0.	1.00	0.
3	19.	18.	1.	0.95	0.05
4	16.	14.	2.	0.88	0.13
5	14.	12.	2.	0.86	0.14
6	19.	19.	0.	1.00	0.
7	15.	15.	0.	1.00	0.
8	6.	6.	0.	1.00	0.
9	2.	2.	0.	1.00	0.
10	4.	3.	1.	0.75	0.25
11	1.	1.	0.	1.00	0.
12	1.	1.	0.	1.00	0.
13	0.	0.	0.	0.	0.
14	0.	0.	0.	0.	0.
15	0.	0.	0.	0.	0.
TOTAL	105.	98.	7.	0.93	0.07

	MEAN	S.D.
CORRECT DOCUMENTS	5.32	2.19
INCORRECT DOCUMENTS	4.57	2.56

TABLE 2 EFFECTIVENESS VS DOCUMENT LENGTH
35 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 2 48 DISCRIMINATING WORDS

SAMPLE DOCUMENTS			CATEGORY A		
NUMBER OF TOKENS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0 - 9	0.	0.	0.	0.	0.
10 - 19	0.	0.	0.	0.	0.
20 - 29	0.	0.	0.	0.	0.
30 - 39	3.	3.	0.	1.00	0.
40 - 49	2.	2.	0.	1.00	0.
50 - 59	6.	5.	1.	0.83	0.17
60 - 69	3.	3.	0.	1.00	0.
70 - 79	4.	4.	0.	1.00	0.
80 - 89	5.	5.	0.	1.00	0.
90 - 99	3.	3.	0.	1.00	0.
100 - 109	0.	0.	0.	0.	0.
110 - 119	2.	2.	0.	1.00	0.
120 - 129	3.	3.	0.	1.00	0.
130 - 139	2.	2.	0.	1.00	0.
140 - 149	1.	1.	0.	1.00	0.
150 - 159	1.	1.	0.	1.00	0.
160 - 169	0.	0.	0.	0.	0.
170 - 179	0.	0.	0.	0.	0.
180 - 189	0.	0.	0.	0.	0.
190 - 199	0.	0.	0.	0.	0.
200 - 209	0.	0.	0.	0.	0.
210 - 219	0.	0.	0.	0.	0.
220 - 229	0.	0.	0.	0.	0.
230 - 239	0.	0.	0.	0.	0.
240 - 249	0.	0.	0.	0.	0.
250 - UP	0.	0.	0.	0.	0.
TOTAL	35.	34.	1.	0.97	0.03

	MEAN	S.D.
CORRECT DOCUMENTS	82.79	31.87
INCORRECT DOCUMENTS	55.00	0.

TABLE 2 EFFECTIVENESS VS DOCUMENT LENGTH
35 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 2 48 DISCRIMINATING WORDS

SAMPLE DOCUMENTS NUMBER OF TOKENS	 NUMBER OF DOCUMENTS	 NUMBER OF CORRECT DOCUMENTS	CATEGORY M NUMBER OF INCORRECT DOCUMENTS	 PERCENTAGE OF CORRECT DOCUMENTS	 PERCENTAGE OF INCORRECT DOCUMENTS
0 - 9	C.	C.	C.	0.	C.
10 - 19	1.	1.	C.	1.00	0.
2C - 29	0.	0.	0.	0.	0.
30 - 39	0.	0.	C.	0.	C.
40 - 49	2.	2.	0.	1.0C	0.
50 - 59	4.	4.	0.	1.C0	0.
60 - 69	3.	1.	2.	0.33	0.67
70 - 79	2.	2.	C.	1.0C	C.
80 - 89	2.	1.	1.	0.50	0.50
90 - 99	3.	3.	C.	1.0C	0.
100 - 109	2.	2.	C.	1.C0	0.
11C - 119	5.	5.	C.	1.0C	0.
120 - 129	4.	4.	C.	1.0C	0.
130 - 139	5.	5.	C.	1.00	0.
140 - 149	2.	1.	1.	0.50	0.50
150 - 159	0.	C.	0.	0.	0.
16C - 169	0.	0.	C.	0.	0.
170 - 179	0.	0.	0.	0.	0.
180 - 189	0.	0.	0.	0.	0.
190 - 199	0.	0.	C.	0.	0.
200 - 209	0.	0.	0.	0.	0.
210 - 219	0.	C.	0.	0.	0.
220 - 229	0.	0.	C.	0.	0.
230 - 239	C.	0.	0.	0.	0.
240 - 249	0.	0.	0.	0.	0.
250 - UP	0.	0.	0.	0.	0.
TOTAL	35.	31.	4.	0.89	0.11

	MEAN	S.D.
CORRECT DOCUMENTS	97.65	34.57
INCORRECT DOCUMENTS	90.50	34.59

TABLE 2 EFFECTIVENESS VS DOCUMENT LENGTH
35 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 2 48 DISCRIMINATING WORDS

SAMPLE DOCUMENTS		CATEGORY P			
NUMBER OF TOKENS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0 - 9	0.	0.	0.	0.	0.
10 - 19	0.	0.	0.	0.	0.
20 - 29	1.	1.	0.	1.00	0.
30 - 39	1.	0.	1.	0.	1.00
40 - 49	1.	1.	0.	1.00	0.
50 - 59	1.	1.	0.	1.00	0.
60 - 69	5.	5.	0.	1.00	0.
70 - 79	1.	1.	0.	1.00	0.
80 - 89	2.	2.	0.	1.00	0.
90 - 99	2.	2.	0.	1.00	0.
100 - 109	4.	4.	0.	1.00	0.
110 - 119	4.	3.	1.	0.75	0.25
120 - 129	1.	1.	0.	1.00	0.
130 - 139	2.	2.	0.	1.00	0.
140 - 149	2.	2.	0.	1.00	0.
150 - 159	3.	3.	0.	1.00	0.
160 - 169	2.	2.	0.	1.00	0.
170 - 179	2.	2.	0.	1.00	0.
180 - 189	0.	0.	0.	0.	0.
190 - 199	0.	0.	0.	0.	0.
200 - 209	0.	0.	0.	0.	0.
210 - 219	0.	0.	0.	0.	0.
220 - 229	0.	0.	0.	0.	0.
230 - 239	0.	0.	0.	0.	0.
240 - 249	0.	0.	0.	0.	0.
250 - UP	1.	1.	0.	1.00	0.
TOTAL	35.	33.	2.	0.94	0.06

	MEAN	S.D.
CORRECT DOCUMENTS	112.27	48.07
INCORRECT DOCUMENTS	76.50	37.50

TABLE 2 EFFECTIVENESS VS DOCUMENT LENGTH
35 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 2 48 DISCRIMINATING WORDS

SAMPLE DOCUMENTS			CATEGORY TOTAL		
NUMBER OF TOKENS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0 - 9	0.	0.	0.	0.	0.
10 - 19	1.	1.	0.	1.00	0.
20 - 29	1.	1.	0.	1.00	0.
30 - 39	4.	3.	1.	0.75	0.25
40 - 49	5.	5.	0.	1.00	0.
50 - 59	11.	10.	1.	0.91	0.09
60 - 69	11.	9.	2.	0.82	0.18
70 - 79	7.	7.	0.	1.00	0.
80 - 89	9.	8.	1.	0.89	0.11
90 - 99	8.	8.	0.	1.00	0.
100 - 109	6.	6.	0.	1.00	0.
110 - 119	11.	10.	1.	0.91	0.09
120 - 129	8.	8.	0.	1.00	0.
130 - 139	9.	9.	0.	1.00	0.
140 - 149	5.	4.	1.	0.80	0.20
150 - 159	4.	4.	0.	1.00	0.
160 - 169	2.	2.	0.	1.00	0.
170 - 179	2.	2.	0.	1.00	0.
180 - 189	0.	0.	0.	0.	0.
190 - 199	0.	0.	0.	0.	0.
200 - 209	0.	0.	0.	0.	0.
210 - 219	0.	0.	0.	0.	0.
220 - 229	0.	0.	[illegible]	0.	0.
230 - 239	0.	0.	[illegible]	0.	0.
240 - 249	0.	0.		0.	0.
250 - UP	1.	1.	0.	1.00	0.
TOTAL	105.	98.	7.	0.93	0.07

	MEAN	S.D.
CORRECT DOCUMENTS	97.42	40.76
INCORRECT DOCUMENTS	81.43	35.20

TABLE 3 EFFECTIVENESS VS WORDS IN INTERSECTION OF DOCUMENTS AND DISCRIMINATING SET
35 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 2 48 DISCRIMINATING WORDS

SAMPLE DOCUMENTS			CATEGORY A		
NUMBER OF WORDS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0	1.	1.	0.	1.00	0.
1	0.	0.	0.	0.	0.
2	3.	2.	1.	0.67	0.33
3	2.	2.	0.	1.00	0.
4	2.	2.	0.	1.00	0.
5	4.	4.	0.	1.00	0.
6	10.	10.	0.	1.00	0.
7	2.	2.	0.	1.00	0.
8	2.	2.	0.	1.00	0.
9	5.	5.	0.	1.00	0.
10	1.	1.	0.	1.00	0.
11	3.	3.	0.	1.00	0.
12	0.	0.	0.	0.	0.
13	0.	0.	0.	0.	0.
14	0.	0.	0.	0.	0.
15	0.	0.	0.	0.	0.
16	0.	0.	0.	0.	0.
17	0.	0.	0.	0.	0.
18	0.	0.	0.	0.	0.
19	0.	0.	0.	0.	0.
20	0.	0.	0.	0.	0.
TOTAL	35.	34.	1.	0.97	0.03

	MEAN	S.D.
CORRECT DOCUMENTS	6.35	2.66
INCORRECT DOCUMENTS	2.00	0.

TABLE 3 EFFECTIVENESS VS WORDS IN INTERSECTION OF DOCUMENTS AND DISCRIMINATING SET
35 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 2 48 DISCRIMINATING WORDS

SAMPLE DOCUMENTS			CATEGORY M		
NUMBER OF WORDS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0	0.	0.	0.	0.	0.
1	1.	1.	0.	1.00	0.
2	1.	1.	0.	1.00	0.
3	2.	1.	1.	0.50	0.50
4	2.	0.	2.	0.	1.00
5	8.	8.	0.	1.00	0.
6	3.	3.	0.	1.00	0.
7	5.	5.	0.	1.00	0.
8	4.	4.	0.	1.00	0.
9	6.	6.	0.	1.00	0.
10	2	1.	1.	0.50	0.50
11	1.	1.	0.	1.00	0.
12	0.	0.	0.	0.	0.
13	0.	0.	0.	0.	0.
14	0.	0.	0.	0.	0.
15	0.	0.	0.	0.	0.
16	0.	0.	0.	0.	0.
17	0.	0.	0.	0.	0.
18	0.	0.	0.	0.	0.
19	0.	0.	0.	0.	0.
20	0.	0.	0.	0.	0.
TOTAL	35.	31.	4.	0.89	0.11

	MEAN	S.D.
CORRECT DOCUMENTS	6.65	2.28
INCORRECT DOCUMENTS	5.25	2.77

TABLE 3 EFFECTIVENESS VS WORDS IN INTERSECTION OF DOCUMENTS AND DISCRIMINATING SET
35 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 2 48 DISCRIMINATING WORDS

SAMPLE DOCUMENTS			CATEGORY P		
NUMBER OF WORDS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0	0.	0.	0.	0.	0.
1	0.	0.	0.	0.	0.
2	0.	0.	0.	0	0.
3	0.	0.	0.	0.	0.
4	7.	5.	2.	0.71	0.29
5	4.	4.	0.	1.00	0.
6	4.	4.	0.	1.00	0.
7	4.	4.	0.	1.00	0.
8	4.	4.	0.	1.00	0.
9	2.	2.	0.	1.00	0.
10	3.	3.	0.	1.00	0.
11	2.	2.	0.	1.00	0.
12	0.	0.	0.	0.	0.
13	1.	1.	0.	1.00	0.
14	3.	3.	0.	1.00	0.
15	0.	0.	0.	0.	0.
16	1.	1.	0.	1.00	0.
17	0.	0.	0.	0.	0.
18	0.	0.	0.	0.	0.
19	0.	0.	0.	0.	0.
20	0.	0.	0.	0.	0.
TOTAL	35.	33.	2.	0.94	0.06

	MEAN	S.D.
CORRECT DOCUMENTS	8.03	3.33
INCORRECT DOCUMENTS	4.00	0.

TABLE 3 EFFECTIVENESS VS WORDS IN INTERSECTION OF DOCUMENTS AND DISCRIMINATING SET
35 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 2 48 DISCRIMINATING WORDS

SAMPLE DOCUMENTS NUMBER OF WORDS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	CATEGORY TOTAL NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0	1.	1.	0.	1.00	0.
1	1.	1.	0.	1.00	0.
2	4.	3.	1.	0.75	0.25
3	4.	3.	1.	0.75	0.25
4	11.	7.	4.	0.64	0.36
5	16.	16.	0.	1.00	0.
6	17.	17.	0.	1.00	0.
7	11.	11.	0.	1.00	0.
8	10.	10.	0.	1.00	0.
9	13.	13.	0.	1.00	0.
10	6.	5.	1.	0.83	0.17
11	6.	6.	0.	1.00	0.
12	0.	0.	0.	0.	0.
13	1.	1.	0.	1.00	0.
14	3.	3.	0.	1.00	0.
15	0.	0.	0.	0.	0.
16	1.	1.	0.	1.00	0.
17	0.	0.	0.	0.	0.
18	0.	0.	0.	0.	0.
19	0.	0.	0.	0.	0.
20	0.	0.	0.	0.	0.
TOTAL	105.	98.	7.	0.93	0.07

	MEAN	S.D.
CORRECT DOCUMENTS	7.01	2.83
INCORRECT DOCUMENTS	4.43	2.34

TABLE 4 EFFECTIVENESS VS RADIUS
35 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 2 48 DISCRIMINATING WORDS

SAMPLE DOCUMENTS			CATEGORY A		
RADIUS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0. - 0.49	12.	11.	1.	0.92	0.08
0.5 - 0.99	22.	22.	C.	1.00	0.
1.0 - 1.99	1.	1.	0.	1.00	0.
2.0 - 2.99	0.	0.	0.	0.	0.
3.0 - 3.99	0.	0.	0.	0	0.
4.0 - 4.99	0.	0.	0.	0.	0.
5.0 - 5.99	0.	0.	0.	0.	0.
6.0 - 6.99	0.	0.	0.	0.	0.
7.0 - 7.99	0.	0.	0.	0.	0.
8.0 - 8.99	0.	0.	0.	0.	0.
9.0 - 9.99	0.	0.	0.	0.	0.
10. - UP	0.	0.	0.	0.	0.
TOTAL	35.	34.	1.	0.97	0.03

	MEAN	S.D.
CORRECT DOCUMENTS	0.61	0.28
INCORRECT DOCUMENTS	0.42	0.

```
TABLE 4    EFFECTIVENESS VS RADIUS
           35 DOCUMENTS IN EACH CATEGORY   DISCRIMINATING WORD SET    2    48 DISCRIMINATING  WORDS

   SAMPLE DOCUMENTS                                 CATEGORY 4
RADIUS        NUMBER OF    NUMBER OF                 NUMBER OF               PERCENTAGE OF            PERCENTAGE OF
              DOCUMENTS      CORRECT DOCUMENTS      INCORRECT DOCUMENTS     CORRECT DOCUMENTS       INCORRECT DOCUMENTS
0.   - 0.49     8.              5.                      3.                    0.63                    0.38

0.5 - 0.99     22.             21.                      1.                    0.95                    0.05

1.0 - 1.99      5.              5.                      0.                    1.00                    0.

2.0 - 2.99      0.              0.                      0.                    0.                      0.

3.0 - 3.99      0.              0.                      0.                    0.                      0.

4.0 - 4.99      0.              0.                      0.                    0.                      0.

5.0 - 5.99      0.              0.                      0.                    0.                      0.

6.0 - 6.99      0.              0.                      0.                    0.                      0.

7.0 - 7.99      0.              0.                      0.                    0.                      0.

8.0 - 8.99      0.              0.                      0.                    0.                      0.

9.0 - 9.99      0.              0.                      0.                    0.                      0.

10.  -    UP    0.              0.                      0.                    0.                      0.

TOTAL          35.             31.                      4.                    0.89                    0.11

                            MEAN              S.D.
CORRECT   DOCUMENTS          0.81              0.25
INCORRECT DOCUMENTS          0.33              0.11
```

TABLE 4 EFFECTIVENESS VS RADIUS
35 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 2 48 DISCRIMINATING WORDS

SAMPLE DOCUMENTS			CATEGORY P		
RADIUS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0. - 0.49	12.	10.	2.	0.83	0.17
0.5 - 0.99	17.	17.	0.	1.00	0.
1.0 - 1.99	6.	6.	0.	1.00	0.
2.0 - 2.99	0.	0.	0.	0.	0.
3.0 - 3.99	0.	0.	0.	0.	0.
4.0 - 4.99	0.	0.	0.	0.	0.
5.0 - 5.99	0.	0.	0.	0.	0.
6.0 - 6.99	0.	0.	0.	0.	0.
7.0 - 7.99	0.	0.	0.	0.	0.
8.0 - 8.99	0.	0.	0.	0.	0.
9.0 - 9.99	0.	0.	0.	0.	0.
10. - UP	0.	0.	0.	0.	0.
TOTAL	35.	33.	2.	0.94	0.06

	MEAN	S.D.
CORRECT DOCUMENTS	0.68	0.30
INCORRECT DOCUMENTS	0.42	0.04

TABLE 4 EFFECTIVENESS VS RADIUS
35 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 2 48 DISCRIMINATING WORDS

SAMPLE DOCUMENTS			CATEGORY TOTAL		
RADIUS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0. - 0.49	32.	26.	6.	0.81	0.19
0.5 - 0.99	61.	60.	1.	0.98	0.02
1.0 - 1.99	12.	12.	0.	1.00	0.
2.0 - 2.99	0.	0.	0.	0.	0.
3.0 - 3.99	0.	0.	0.	0.	0.
4.0 - 4.99	0.	0.	0.	0.	0.
5.0 - 5.99	0.	0.	0.	0.	0.
6.0 - 6.99	0.	0.	0.	0.	0.
7.0 - 7.99	0.	0.	0.	0.	0.
8.0 - 8.99	0.	0.	0.	0.	0.
9.0 - 9.99	0.	0.	0.	0.	0.
10. - UP	0.	0.	0.	0.	0.
TOTAL	105.	98.	7.	0.93	0.07

	MEAN	S.D.
CORRECT DOCUMENTS	0.70	0.29
INCORRECT DOCUMENTS	0.37	0.09

```
TABLE 5   DOCUMENT CLASSIFICATION SUMMARY
          35 DOCUMENTS IN EACH CATEGORY   DISCRIMINATING WORD SET    2      48 DISCRIMINATING  WORDS

SAMPLE
                  AUTO   CATEGORY
ACTUAL
CATEGORY       A         M         P         TOTAL

A             34.00      1.00      0.        35.00

M              1.00     31.00      3.00      35.00

P              0.        2.00     33.00      35.00

TOTAL         35.00     34.00     36.00     105.00

                   PERCENTAGE

A              0.97      0.03      0.         1.00

M              0.03      0.89      0.09       1.00

P              0.        0.06      0.94       1.00

                     SWETS MEASURES              RECALL       RELEVANCE      PRECISION
CATEGORY A        PERTINENT   NOT PERTINENT      RATIO          RATIO          RATIO
 RETRIEVED          0.97          0.01            0.97           0.97           1.00


                     SWETS MEASURES              RECALL       RELEVANCE      PRECISION
CATEGORY M        PERTINENT   NOT PERTINENT      RATIO          RATIO          RATIO
 RETRIEVED          0.89          0.04            0.89           0.91           0.97


                     SWETS MEASURES              RECALL       RELEVANCE      PRECISION
CATEGORY P        PERTINENT   NOT PERTINENT      RATIO          RATIO          RATIO
 RETRIEVED          0.94          0.04            0.94           0.92           1.03
```

TABLE 1 EFFECTIVENESS VS NUMBER OF SENTENCES
35 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 2 48 DISCRIMINATING WORDS

TEST DOCUMENTS			CATEGORY A		
NUMBER OF SENTENCES	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
1	5.	5.	0.	1.00	0.
2	40.	32.	8.	0.80	0.20
3	66.	54.	12.	0.82	0.18
4	64.	46.	18.	0.72	0.28
5	67.	56.	11.	0.84	0.16
6	36.	26.	10.	0.72	0.28
7	26.	20.	6.	0.77	0.23
8	13.	11.	2.	0.85	0.15
9	3.	2.	1.	0.67	0.33
10	3.	2.	1.	0.67	0.33
11	1.	1.	0.	1.00	0.
12	1.	1.	0.	1.00	0.
13	1.	1.	0.	1.00	0.
14	0.	0.	0.	0.	0.
15	0.	0.	0.	0.	0.
TOTAL	326.	257.	69.	0.79	0.21

	MEAN	S.D.
CORRECT DOCUMENTS	4.49	1.97
INCORRECT DOCUMENTS	4.58	1.79

TABLE 1 EFFECTIVENESS VS NUMBER OF SENTENCES
35 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 2 48 DISCRIMINATING WORDS

TEST DOCUMENTS		CATEGORY M			
NUMBER OF SENTENCES	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
1	11.	7.	4.	0.64	0.36
2	33.	14.	19.	0.42	0.58
3	54.	33.	21.	0.61	0.39
4	71.	44.	27.	0.62	0.38
5	58.	38.	20.	0.66	0.34
6	46.	31.	15.	0.67	0.33
7	23.	16.	7.	0.70	0.30
8	26.	19.	7.	0.73	0.27
9	18.	15.	3.	0.83	0.17
10	4.	4.	0.	1.00	0.
11	3.	2.	1.	0.67	0.33
12	1.	1.	0.	1.00	0.
13	0.	0.	0.	0.	0.
14	2.	1.	1.	0.50	0.50
15	0.	0.	0.	0.	0.
TOTAL	350.	225.	125.	0.64	0.36

	MEAN	S.D.
CORRECT DOCUMENTS	5.21	2.32
INCORRECT DOCUMENTS	4.48	2.16

TABLE 1 EFFECTIVENESS VS NUMBER OF SENTENCES
35 DOCUMENTS IN EACH CATEGORY DISCRIMINATING *ORD SET 2 48 DISCRIMINATING *ORDS

TEST DOCUMENTS			CATEGORY P		
NUMBER OF SENTENCES	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
1	12.	6.	6.	0.50	0.50
2	69.	46.	23.	0.67	0.33
3	150.	93.	57.	0.62	0.38
4	179.	127.	52.	0.71	0.29
5	165.	108.	57.	0.65	0.35
6	135.	91.	44.	0.67	0.33
7	98.	67.	31.	0.68	0.32
8	68.	55.	13.	0.81	0.19
9	49.	39.	10.	0.80	0.20
10	25.	18.	7.	0.72	0.28
11	11.	8.	3.	0.73	0.27
12	7.	3.	4.	0.43	0.57
13	1.	0.	1.	0.	1.00
14	2.	1.	1.	0.50	0.50
15	0.	0.	0.	0.	0.
TOTAL	971.	662.	309.	0.68	0.32

	MEAN	S.D.
CORRECT DOCUMENTS	5.36	2.26
INCORRECT DOCUMENTS	5.08	2.31

TABLE 1 EFFECTIVENESS VS NUMBER OF SENTENCES
35 DOCUMENTS IN EACH CATFGORY DISCRIMINATING WORD SET 2 48 DISCRIMINATING WORDS

TEST DOCUMENTS		CATEGORY TOTAL			
NUMBER OF SENTENCES	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
1	28.	18.	10.	0.64	0.36
2	142.	92.	50.	0.65	0.35
3	270.	180.	90.	0.67	0.33
4	314.	217.	97.	0.69	0.31
5	290.	202.	88.	0.70	0.30
6	217.	148.	69.	0.68	0.32
7	147.	103.	44.	0.70	0.30
8	107.	85.	22.	0.79	0.21
9	70.	56.	14.	0.80	0.20
10	32.	24.	8.	0.75	0.25
11	15.	11.	4.	0.73	0.27
12	9.	5.	4.	0.56	0.44
13	2.	1.	1.	0.50	0.50
14	4.	2.	2.	0.50	0.50
15	0.	0.	0.	0.	0.
TOTAL	1647.	1144.	503.	0.69	0.31

	MEAN	S.D.
CORRECT DOCUMENTS	5.14	2.24
INCORRECT DOCUMENTS	4.86	2.23

TABLE 2 EFFECTIVENESS VS DOCUMENT LENGTH
35 DOCUMENTS IN EACH CATEGORY   DISCRIMINATING WORD SET 2   48 DISCRIMINATING WORDS

TEST DOCUMENTS			CATEGORY A		
NUMBER OF TOKENS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0 - 9	0.	0.	0.	0.	0.
10 - 19	0.	0.	0.	0.	0.
20 - 29	4.	4.	0.	1.00	0.
30 - 39	15.	14.	1.	0.93	0.07
40 - 49	38.	28.	10.	0.74	0.26
50 - 59	33.	25.	8.	0.76	0.24
60 - 69	42.	36.	6.	0.86	0.14
70 - 79	42.	37.	5.	0.88	0.12
80 - 89	36.	24.	12.	0.67	0.33
90 - 99	34.	21.	13.	0.62	0.38
100 - 109	29.	23.	6.	0.79	0.21
110 - 119	15.	13.	2.	0.87	0.13
120 - 129	17.	15.	2.	0.88	0.12
130 - 139	4.	4.	0.	1.00	0.
140 - 149	4.	4.	0.	1.00	0.
150 - 159	4.	2.	2.	0.50	0.50
160 - 169	4.	2.	2.	0.50	0.50
170 - 179	2.	2.	0.	1.00	0.
180 - 189	3.	3.	0.	1.00	0.
190 - 199	0.	0.	0.	0.	0.
200 - 209	0.	0.	0.	0.	0.
210 - 219	0.	0.	0.	0.	0.
220 - 229	0.	0.	0.	0.	0.
230 - 239	0.	0.	0.	0.	0.
240 - 249	0.	0.	0.	0.	0.
250 - UP	0.	0.	0.	0.	0.
TOTAL	326.	257.	69.	0.79	0.21

	MEAN	S.D.
CORRECT DOCUMENTS	80.49	32.39
INCORRECT DOCUMENTS	82.77	29.63

TABLE 2   EFFECTIVENESS VS DOCUMENT LENGTH
35 DOCUMENTS IN EACH CATEGORY   DISCRIMINATING WORD SET   2   48 DISCRIMINATING WORDS

TEST DOCUMENTS			CATEGORY M		
NUMBER OF TOKENS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0 - 9	0.	0.	0.	0.	0.
10 - 19	1.	0.	1.	0.	1.00
20 - 29	8.	2.	6.	0.25	0.75
30 - 39	18.	11.	7.	0.61	0.39
40 - 49	23.	14.	9.	0.61	0.39
50 - 59	33.	19.	14.	0.58	0.42
60 - 69	40.	22.	18.	0.55	0.45
70 - 79	32.	16.	16.	0.50	0.50
80 - 89	38.	30.	8.	0.79	0.21
90 - 99	25.	14.	11.	0.56	0.44
100 - 109	26.	19.	7.	0.73	0.27
110 - 119	19.	12.	7.	0.63	0.37
120 - 129	25.	20.	5.	0.80	0.20
130 - 139	12.	11.	1.	0.92	0.08
140 - 149	9.	8.	1.	0.89	0.11
150 - 159	11.	6.	5.	0.55	0.45
160 - 169	5.	5.	0.	1.00	0.
170 - 179	9.	5.	4.	0.56	0.44
180 - 189	7.	5.	2.	0.71	0.29
190 - 199	2.	2.	0.	1.00	0.
200 - 209	4.	1.	3.	0.25	0.75
210 - 219	1.	1.	0.	1.00	0.
220 - 229	1.	1.	0.	1.00	0.
230 - 239	0.	0.	0.	0.	0.
240 - 249	1.	1.	0.	1.00	0.
250 - UP	0.	0.	0.	0.	0.
TOTAL	350.	225.	125.	0.64	0.36

	MEAN	S.D.
CORRECT DOCUMENTS	97.15	42.55
INCORRECT DOCUMENTS	84.13	41.55

TABLE 2 EFFECTIVENESS VS DOCUMENT LENGTH
35 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 2 48 DISCRIMINATING WORDS

TEST DOCUMENTS			CATEGORY P		
NUMBER OF TOKENS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0 - 9	0.	0.	0.	0.	0.
10 - 19	0.	0.	0.	0.	0.
20 - 29	8.	3.	5.	0.38	0.63
30 - 39	34.	22.	12.	0.65	0.35
40 - 49	50.	32.	18.	0.64	0.36
50 - 59	74.	48.	26.	0.65	0.35
60 - 69	98.	60.	38.	0.61	0.39
70 - 79	92.	60.	32.	0.65	0.35
80 - 89	77.	52.	25.	0.68	0.32
90 - 99	71.	44.	27.	0.62	0.38
100 - 109	66.	44.	22.	0.67	0.33
110 - 119	61.	40.	21.	0.66	0.34
120 - 129	77.	59.	18.	0.77	0.23
130 - 139	58.	39.	19.	0.67	0.33
140 - 149	47.	36.	11.	0.77	0.23
150 - 159	32.	26.	6.	0.81	0.19
160 - 169	30.	23.	7.	0.77	0.23
170 - 179	29.	22.	7.	0.76	0.24
180 - 189	21.	19.	2.	0.90	0.10
190 - 199	16.	14.	2.	0.88	0.13
200 - 209	11.	7.	4.	0.64	0.36
210 - 219	4.	2.	2.	0.50	0.50
220 - 229	4.	3.	1.	0.75	0.25
230 - 239	9.	6.	3.	0.67	0.33
240 - 249	1.	1.	0.	1.00	0.
250 - UP	1.	0.	1.	0.	1.00
TOTAL	971.	662.	309.	0.68	0.32

	MEAN	S.D.
CORRECT DOCUMENTS	106.35	46.00
INCORRECT DOCUMENTS	96.59	44.20

TABLE 2 EFFECTIVENESS VS DOCUMENT LENGTH
35 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 2 48 DISCRIMINATING WORDS

TEST DOCUMENTS			CATEGORY TOTAL		
NUMBER OF TOKENS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0 - 9	0.	0.	0.	0.	0.
10 - 19	1.	0.	1.	0.	1.00
20 - 29	20.	9.	11.	0.45	0.55
30 - 39	67.	47.	20.	0.70	0.30
40 - 49	111.	74.	37.	0.67	0.33
50 - 59	140.	92.	48.	0.66	0.34
60 - 69	180.	118.	62.	0.66	0.34
70 - 79	166.	113.	53.	0.68	0.32
80 - 89	151.	106.	45.	0.70	0.30
90 - 99	130.	79.	51.	0.61	0.39
100 - 109	121.	86.	35.	0.71	0.29
110 - 119	95.	65.	30.	0.68	0.32
120 - 129	119.	94.	25.	0.79	0.21
130 - 139	74.	54.	20.	0.73	0.27
140 - 149	60.	48.	12.	0.80	0.20
150 - 159	47.	34.	13.	0.72	0.28
160 - 169	39.	30.	9.	0.77	0.23
170 - 179	40.	29.	11.	0.72	0.27
180 - 189	31.	27.	4.	0.87	0.13
190 - 199	18.	16.	2.	0.89	0.11
200 - 209	15.	8.	7.	0.53	0.47
210 - 219	5.	3.	2.	0.60	0.40
220 - 229	5.	4.	1.	0.80	0.20
230 - 239	9.	6.	3.	0.67	0.33
240 - 249	2.	2.	0.	1.00	0.
250 - UP	1.	0.	1.	0.	1.00
TOTAL	1647.	1144.	503.	0.69	0.31

	MEAN	S.D.
CORRECT DOCUMENTS	99.08	43.94
INCORRECT DOCUMENTS	91.60	42.30

TABLE 3 EFFECTIVENESS VS WORDS IN INTERSECTION OF DOCUMENTS AND DISCRIMINATING SET
35 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 2 48 DISCRIMINATING WORDS

TEST DOCUMENTS			CATEGORY A		
NUMBER OF WORDS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0	0.	0.	0.	0.	0.
1	1.	1.	0.	1.00	0.
2	6.	4.	2.	0.67	0.33
3	24.	18.	6.	0.75	0.25
4	40.	29.	11.	0.72	0.27
5	45.	35.	10.	0.78	0.22
6	47.	38.	9.	0.81	0.19
7	44.	37.	7.	0.84	0.16
8	45.	40.	5.	0.89	0.11
9	27.	21.	6.	0.78	0.22
10	21.	14.	7.	0.67	0.33
11	13.	10.	3.	0.77	0.23
12	7.	6.	1.	0.86	0.14
13	2.	2.	0.	1.00	0.
14	4.	2.	2.	0.50	0.50
15	0.	0.	0.	0.	0.
16	0.	0.	0.	0.	0.
17	0.	0.	0.	0.	0.
18	0.	0.	0.	0.	0.
19	0.	0.	0.	0.	0.
20	0.	0.	0.	0.	0.
TOTAL	326.	257.	69.	0.79	0.21

	MEAN	S.D.
CORRECT DOCUMENTS	6.72	2.47
INCORRECT DOCUMENTS	6.61	2.81

TABLE 3 EFFECTIVENESS VS WORDS IN INTERSECTION OF DOCUMENTS AND DISCRIMINATING SET
35 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 2 48 DISCRIMINATING WORDS

TEST DOCUMENTS		CATEGORY M			
NUMBER OF WORDS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0	6.	0.	6.	0.	1.00
1	5.	3.	2.	0.60	0.40
2	20.	9.	11.	0.45	0.55
3	29.	11.	18.	0.38	0.62
4	33.	18.	15.	0.55	0.45
5	42.	25.	17.	0.60	0.40
6	54.	36.	18.	0.67	0.33
7	48.	34.	14.	0.71	0.29
8	48.	35.	13.	0.73	0.27
9	26.	19.	7.	0.73	0.27
10	11.	9.	2.	0.82	0.18
11	11.	10.	1.	0.91	0.09
12	9.	8.	1.	0.89	0.11
13	5.	5.	0.	1.00	0.
14	0.	0.	0.	0.	0.
15	0.	0.	0.	0.	0.
16	2.	2.	0.	1.00	0.
17	0.	0.	0.	0.	0.
18	0.	0.	0.	0.	0.
19	1.	1.	0.	1.00	0.
20	0.	0.	0.	0.	0.
TOTAL	350.	225.	125.	0.64	0.36

	MEAN	S.D.
CORRECT DOCUMENTS	6.97	2.86
INCORRECT DOCUMENTS	5.11	2.54

TABLE 3 EFFECTIVENESS VS WORDS IN INTERSECTION OF DOCUMENTS AND DISCRIMINATING SET
35 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 2 48 DISCRIMINATING WORDS

TEST DOCUMENTS		CATEGORY P			
NUMBER OF WORDS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0	2.	0.	2.	0.	1.00
1	11.	7.	4.	0.64	0.36
2	22.	11.	11.	0.50	0.50
3	41.	21.	20.	0.51	0.49
4	74.	47.	27.	0.64	0.36
5	106.	62.	44.	0.58	0.42
6	116.	78.	38.	0.67	0.33
7	138.	88.	50.	0.64	0.36
8	102.	72.	30.	0.71	0.29
9	106.	80.	26.	0.75	0.25
10	77.	60.	17.	0.78	0.22
11	50.	35.	15.	0.70	0.30
12	52.	42.	10.	0.81	0.19
13	35.	27.	8.	0.77	0.23
14	17.	12.	5.	0.71	0.29
15	10.	9.	1.	0.90	0.10
16	4.	4.	0.	1.00	0.
17	5.	4.	1.	0.80	0.20
18	2.	2.	0.	1.00	0.
19	1.	1.	0.	1.00	0.
20	0.	0.	0.	0.	0.
TOTAL	971.	662.	309.	0.68	0.32

	MEAN	S.D.
CORRECT DOCUMENTS	8.61	3.19
INCORRECT DOCUMENTS	6.88	2.99

TABLE 3   EFFECTIVENESS VS WORDS IN INTERSECTION OF DOCUMENTS AND DISCRIMINATING SET
35 DOCUMENTS IN EACH CATEGORY   DISCRIMINATING WORD SET 2   48 DISCRIMINATING WORDS

TEST DOCUMENTS			CATEGORY TOTAL		
NUMBER OF WORDS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0	8.	0.	8.	0.	1.00
1	17.	11.	6.	0.65	0.35
2	48.	24.	24.	0.50	0.50
3	94.	50.	44.	0.53	0.47
4	147.	94.	53.	0.64	0.36
5	193.	122.	71.	0.63	0.37
6	217.	152.	65.	0.70	0.30
7	230.	159.	71.	0.69	0.31
8	195.	147.	48.	0.75	0.25
9	159.	120.	39.	0.75	0.25
10	109.	83.	26.	0.76	0.24
11	74.	55.	19.	0.74	0.26
12	68.	56.	12.	0.82	0.13
13	42.	34.	8.	0.81	0.19
14	21.	14.	7.	0.67	0.33
15	10.	9.	1.	0.90	0.10
16	6.	6.	0.	1.00	0.
17	5.	4.	1.	0.80	0.20
18	2.	2.	0.	1.00	0.
19	2.	2.	0.	1.00	0.
20	0.	0.	0.	0.	0.
TOTAL	1647.	1144.	503.	0.69	0.31

	MEAN	S.D.
CORRECT DOCUMENTS	7.51	3.04
INCORRECT DOCUMENTS	6.41	2.34

TABLE 4 EFFECTIVENESS VS RADIUS
35 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 2 48 DISCRIMINATING WORDS

TEST DOCUMENTS			CATEGORY A		
RADIUS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0. - 0.49	111.	79.	32.	0.71	0.29
0.5 - 0.99	110.	85.	25.	0.77	0.23
1.0 - 1.99	90.	79.	11.	0.88	0.12
2.0 - 2.99	12.	11.	1.	0.92	0.08
3.0 - 3.99	2.	2.	0.	1.00	0.
4.0 - 4.99	1.	1.	0.	1.00	0.
5.0 - 5.99	0.	0.	0.	0.	0.
6.0 - 6.99	0.	0.	0.	0.	0.
7.0 - 7.99	0.	0.	0.	0.	0.
8.0 - 8.99	0.	0.	0.	0.	0.
9.0 - 9.99	0.	0.	0.	0.	0.
10. - UP	0.	0.	0.	0.	0.
TOTAL	326.	257.	69.	0.79	0.21

	MEAN	S.D.
CORRECT DOCUMENTS	0.91	0.65
INCORRECT DOCUMENTS	0.65	0.45

TABLE 4 EFFECTIVENESS VS RADIUS
35 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 2 48 DISCRIMINATING WORDS

TEST DOCUMENTS			CATEGORY M		
RADIUS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0. - 0.49	135.	55.	80.	0.41	0.59
0.5 - 0.99	123.	86.	37.	0.70	0.30
1.0 - 1.99	80.	72.	8.	0.90	0.10
2.0 - 2.99	12.	12.	0.	1.00	0.
3.0 - 3.99	0.	0.	0.	0.	0.
4.0 - 4.99	0.	0.	0.	0.	0.
5.0 - 5.99	0.	0.	0.	0.	0.
6.0 - 6.99	0.	0.	0.	0.	0.
7.0 - 7.99	0.	0.	0.	0.	0.
8.0 - 8.99	0.	0.	0.	0.	0.
9.0 - 9.99	0.	0.	0.	0.	0.
10. - UP	0.	0.	0.	0.	0.
TOTAL	350.	225.	125.	0.64	0.36

	MEAN	S.D.
CORRECT DOCUMENTS	0.92	0.54
INCORRECT DOCUMENTS	0.43	0.35

```
TABLE 4    EFFECTIVENESS VS RADIUS
           35 DOCUMENTS IN EACH CATEGORY   DISCRIMINATING WORD SET    2     48 DISCRIMINATING  WORDS

      TEST DOCUMENTS                                 CATEGORY P
RADIUS        NUMBER OF    NUMBER OF                   NUMBER OF                 PERCENTAGE OF           PERCENTAGE OF
              DOCUMENTS     CORRECT DOCUMENTS       INCORRECT DOCUMENTS       CORRECT DOCUMENTS       INCORRECT DOCUMENTS
C.   - 0.49   419.            251.                     168.                       0.60                    0.40

0.5 - 0.99    358.            253.                     105.                       0.71                    0.29

1.0 - 1.99    179.            145.                      34.                       0.81                    0.19

2.0 - 2.99     14.             12.                       2.                       0.86                    0.14

3.0 - 3.99      1.              1.                       0.                       1.00                    0.

4.0 - 4.99      0.              0.                       0.                       0.                      0.

5.0 - 5.99      0.              0.                       C.                       0.                      0.

6.0 - 6.99      0.              0.                       0.                       0.                      0.

7.0 - 7.99      0.              0.                       0.                       0.                      0.

8.0 - 8.99      0.              0.                       0.                       0.                      0.

9.0 - 9.99      0.              0.                       0.                       0.                      0.

10. -     UP    0.              0.                       0.                       0.                      0.

TOTAL         971.            662.                     309.                       0.68                    0.32

                            MEAN               S.D.
CORRECT   DOCUMENTS          0.72               0.48
INCORRECT DOCUMENTS          0.57               0.39
```

TABLE 4 EFFECTIVENESS VS RADIUS
35 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 2 48 DISCRIMINATING WORDS

TEST DOCUMENTS RADIUS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	CATEGORY TOTAL NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0. - 0.49	665.	385.	280.	0.58	0.42
0.5 - 0.99	591.	424.	167.	0.72	0.28
1.0 - 1.99	349.	296.	53.	0.85	0.15
2.0 - 2.99	38.	35.	3.	0.92	0.08
3.0 - 3.99	3.	3.	0.	1.00	0.
4.0 - 4.99	1.	1.	0.	1.09	0.
5.0 - 5.99	0.	0.	.	0.	0.
6.0 - 6.99	0.	0.	0	0.	0.
7.0 - 7.99	0.	0.	0.	0.	0.
8.0 - 8.99	0.	0.	0.	0.	0.
9.0 - 9.99	0.	0.	0.	0.	0.
10. - UP	0.	0.	0.	0.	0.
TOTAL	1647.	1144.	503.	0.69	0.31

	MEAN	S.D.
CORRECT DOCUMENTS	0.80	0.54
INCORRECT DOCUMENTS	0.54	0.40

```
TABLE 5   DOCUMENT CLASSIFICATION SUMMARY
          35 DOCUMENTS IN EACH CATEGORY   DISCRIMINATING WORD SET    2      48 DISCRIMINATING  WORDS

  TEST
                    AUTO    CATEGORY
ACTUAL
CATEGORY         A         M         P        TOTAL

A             257.00     15.00     54.00    326.00

M              45.00    225.00     80.00    350.00

P             105.00    204.00    662.00    971.00

TOTAL         407.00    444.00    796.00   1647.00

                     PERCENTAGE

A               0.79      0.05      0.17      1.00

M               0.13      0.64      0.23      1.00

P               0.11      0.21      0.68      1.00

                         SWETS MEASURES                 RECALL        RELEVANCE        PRECISION
CATEGORY A            PERTINENT   NOT PERTINENT          RATIO           RATIO            RATIO
 RETRIEVED              0.79          0.10                0.79            0.53             1.25


                         SWETS MEASURES                 RECALL        RELEVANCE        PRECISION
CATEGORY M            PERTINENT   NOT PERTINENT          RATIO           RATIO            RATIO
 RETRIEVED              0.64          0.16                0.64            0.51             1.27


                         SWETS MEASURES                 RECALL        RELEVANCE        PRECISION
CATEGORY P            PERTINENT   NOT PERTINENT          RATIO           RATIO            RATIO
 RETRIEVED              0.68          0.17                0.68            0.83             0.82
```

TABLE 1 EFFECTIVENESS VS NUMBER OF SENTENCES
70 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 2 48 DISCRIMINATING WORDS

SAMPLE DOCUMENTS			CATEGORY A		
NUMBER OF SENTENCES	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
1	1.	1.	0.	1.00	0.
2	5.	4.	1.	0.80	0.20
3	19.	16.	3.	0.84	0.16
4	16.	16.	0.	1.00	0.
5	9.	9.	0.	1.00	0.
6	9.	7.	2.	0.78	0.22
7	5.	5.	0.	1.00	0.
8	5.	5.	0.	1.00	0.
9	0.	0.	0.	0.	0.
10	0.	0.	0.	0.	0.
11	0.	0.	0.	0.	0.
12	1.	1.	0.	1.00	0.
13	0.	0.	0.	0.	0.
14	0.	0.	0.	0.	0.
15	0.	0.	0.	0.	0.
TOTAL	70.	64.	6.	0.91	0.09

	MEAN	S.D.
CORRECT DOCUMENTS	4.61	1.95
INCORRECT DOCUMENTS	3.83	1.57

TABLE 1 EFFECTIVENESS VS NUMBER OF SENTENCES
70 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 2 48 DISCRIMINATING WORDS

SAMPLE DOCUMENTS			CATEGORY M		
NUMBER OF SENTENCES	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
1	1.	1.	0.	1.00	0.
2	7.	5.	2.	0.71	0.29
3	9.	9.	0.	1.00	0.
4	16.	11.	5.	0.69	0.31
5	9.	7.	2.	0.78	0.22
6	9.	7.	2.	0.78	0.22
7	7.	6.	1.	0.86	0.14
8	6.	5.	1.	0.83	0.17
9	6.	6.	0.	1.00	0.
10	0.	0.	0.	0.	0.
11	0.	0.	0.	0.	0.
12	0.	0.	0.	0.	0.
13	0.	0.	0.	0.	0.
14	0.	0.	0.	0.	0.
15	0.	0.	0.	0.	0.
TOTAL	70.	57.	13.	0.81	0.19

	MEAN	S.D.
CORRECT DOCUMENTS	5.18	2.21
INCORRECT DOCUMENTS	4.69	1.68

TABLE 1 EFFECTIVENESS VS NUMBER OF SENTENCES
70 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 2 48 DISCRIMINATING WORDS

SAMPLE DOCUMENTS			CATEGORY P		
NUMBER OF SENTENCES	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
1	0.	0.	0.	0.	0.
2	6.	5.	1.	0.83	0.17
3	16.	9.	7.	0.56	0.44
4	11.	7.	4.	0.64	0.36
5	12.	11.	1.	0.92	0.08
6	8.	8.	0.	1.00	0.
7	7.	5.	2.	0.71	0.29
8	3.	3.	0.	1.00	0.
9	2.	2.	0.	1.00	0.
10	2.	2.	0.	1.00	0.
11	2.	2.	0.	1.00	0.
12	0.	0.	0.	0.	0.
13	0.	0.	0.	0.	0.
14	0.	0.	0.	0.	0.
15	0.	0.	0.	0.	0.
TOTAL	69.	54.	15.	0.78	0.22

	MEAN	S.D.
CORRECT DOCUMENTS	5.31	2.32
INCORRECT DOCUMENTS	3.87	1.41

TABLE 1 EFFECTIVENESS VS NUMBER OF SENTENCES
70 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 2 48 DISCRIMINATING WORDS

SAMPLE DOCUMENTS			CATEGORY TOTAL		
NUMBER OF SENTENCES	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
1	2.	2.	0.	1.00	0.
2	18.	14.	4.	0.78	0.22
3	44.	34.	10.	0.77	0.23
4	43.	34.	9.	0.79	0.21
5	30.	27.	3.	0.90	0.10
6	26.	22.	4.	0.85	0.15
7	19.	16.	3.	0.84	0.16
8	14.	13.	1.	0.93	0.07
9	8.	8.	0.	1.00	0.
10	2.	2.	0.	1.00	0.
11	2.	2.	0.	1.00	0.
12	1.	1.	0.	1.00	0.
13	0.	0.	0.	0.	0.
14	0.	0.	0.	0.	0.
15	0.	0.	0.	0.	0.
TOTAL	209.	175.	34.	0.84	0.16

	MEAN	S.D.
CORRECT DOCUMENTS	5.01	2.18
INCORRECT DOCUMENTS	4.18	1.60

TABLE 2 EFFECTIVENESS VS DOCUMENT LENGTH
70 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 2 48 DISCRIMINATING WORDS

SAMPLE DOCUMENTS			CATEGORY A		
NUMBER OF TOKENS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0 - 9	0.	0.	0.	0.	0.
10 - 19	0.	0.	0.	0.	0.
20 - 29	0.	0.	0.	0.	0.
30 - 39	3.	2.	1.	0.67	0.33
40 - 49	13.	13.	0.	1.00	0.
50 - 59	7.	5.	2.	0.71	0.29
60 - 69	6.	5.	1.	0.83	0.17
70 - 79	10.	10.	0.	1.00	0.
80 - 89	9.	9.	0.	1.00	0.
90 - 99	7.	6.	1.	0.86	0.14
100 - 109	4.	4.	0.	1.00	0.
110 - 119	3.	3.	0.	1.00	0.
120 - 129	1.	1.	0.	1.00	0.
130 - 139	1.	0.	1.	0.	1.00
140 - 149	0.	0.	0.	0.	0.
150 - 159	3.	3.	0.	1.00	0.
160 - 169	0.	0.	0.	0.	0.
170 - 179	1.	1.	0.	1.00	0.
180 - 189	2.	2.	0.	1.00	0.
190 - 199	0.	0.	0.	0.	0.
200 - 209	0.	0.	0.	0.	0.
210 - 219	0.	0.	0.	0.	0.
220 - 229	0.	0.	0.	0.	0.
230 - 239	0.	0.	0.	0.	0.
240 - 249	0.	0.	0.	0.	0.
250 - UP	0.	0.	0.	0.	0.
TOTAL	70.	64.	6.	0.91	0.09

	MEAN	S.D.
CORRECT DOCUMENTS	81.67	35.51
INCORRECT DOCUMENTS	72.83	32.34

TABLE 2 EFFECTIVENESS VS DOCUMENT LENGTH
70 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 2 48 DISCRIMINATING WORDS

SAMPLE- DOCUMENTS		CATEGORY M			
NUMBER OF TOKENS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0 - 9	0.	0.	0.	0.	0.
10 - 19	0.	0.	0.	0.	0.
20 - 29	1.	1.	0.	1.00	0.
30 - 39	4.	3.	1.	0.75	0.25
40 - 49	2.	1.	1.	0.50	0.50
50 - 59	12.	11.	1.	0.92	0.08
60 - 69	6.	4.	2.	0.67	0.33
70 - 79	5.	3.	2.	0.60	0.40
80 - 89	8.	8.	0.	1.00	0.
90 - 99	4.	2.	2.	0.50	0.50
100 - 109	4.	3.	1.	0.75	0.25
110 - 119	4.	3.	1.	0.75	0.25
120 - 129	5.	5.	0.	1.00	0.
130 - 139	4.	3.	1.	0.75	0.25
140 - 149	1.	1.	0.	1.00	0.
150 - 159	2.	1.	1.	0.50	0.50
160 - 169	2.	2.	0.	1.00	0.
170 - 179	2.	2.	0.	1.00	0.
180 - 189	2.	2.	0.	1.00	0.
190 - 199	1.	1.	0.	1.00	0.
200 - 209	0.	0.	0.	0.	0.
210 - 219	1.	1.	0.	1.00	0.
220 - 229	0.	0.	0.	0.	0.
230 - 239	0.	0.	0.	0.	0.
240 - 249	0.	0.	0.	0.	0.
250 - UP	0.	0.	0.	0.	0.
TOTAL	70.	57.	13.	0.81	0.19

	MEAN	S.D.
CORRECT DOCUMENTS	97.72	45.52
INCORRECT DOCUMENTS	86.52	33.48

TABLE 2 EFFECTIVENESS VS DOCUMENT LENGTH
70 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 2 48 DISCRIMINATING WORDS

SAMPLE DOCUMENTS		CATEGORY P			
NUMBER OF TOKENS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0 - 9	0.	0.	C.	0.	0.
10 - 19	0.	0.	0.	0.	0.
20 - 29	0.	0.	0.	0.	0.
30 - 39	1.	1.	0.	1.00	0.
40 - 49	7.	4.	3.	0.57	0.43
50 - 59	7.	4.	3.	0.57	0.43
60 - 69	7.	5.	2.	0.71	0.29
70 - 79	7.	5.	2.	0.71	0.29
80 - 89	2.	2.	C.	1.0C	0.
90 - 99	4.	3.	1.	0.75	0.25
100 - 109	7.	6.	1.	C.86	0.14
110 - 119	3.	2.	1.	0.67	0.33
120 - 129	3.	3.	0.	1.00	0.
130 - 139	4.	2.	2.	0.5C	0.50
140 - 149	5.	5.	0.	1.0C	0.
150 - 159	2.	2.	C.	1.00	0.
160 - 169	7.	7.	0.	1.00	0.
170 - 179	0.	0.	0.	0.	0.
180 - 189	1.	1.	0.	1.0C	0.
190 - 199	0.	C.	0.	0.	0.
200 - 209	0.	0.	0.	0.	C.
210 - 219	0.	0.	0.	0.	0.
220 - 229	0.	0.	0.	0.	0.
230 - 239	2.	2.	0.	1.00	0.
240 - 249	0.	0.	0.	0.	0.
250 - UP	0.	0.	0.	0.	0.
TOTAL	69.	54.	15.	0.78	0.22

	MEAN	S.D.
CORRECT DOCUMENTS	109.37	47.56
INCORRECT DOCUMENTS	76.47	29.80

TABLE 2 EFFECTIVENESS VS DOCUMENT LENGTH
70 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 2 48 DISCRIMINATING WORDS

SAMPLE DOCUMENTS			CATEGORY TOTAL		
NUMBER OF TOKENS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0 - 9	0.	0.	0.	0.	0.
10 - 19	0.	0.	0.	0.	0.
20 - 29	1.	1.	0.	1.00	0.
30 - 39	8.	6.	2.	0.75	0.25
40 - 49	22.	18.	4.	0.82	0.18
50 - 59	26.	20.	6.	0.77	0.23
60 - 69	19.	14.	5.	0.74	0.26
70 - 79	22.	18.	4.	0.82	0.18
80 - 89	19.	19.	0.	1.00	0.
90 - 99	15.	11.	4.	0.73	0.27
100 - 109	15.	13.	2.	0.87	0.13
110 - 119	10.	8.	2.	0.80	0.20
120 - 129	9.	9.	0.	1.00	0.
130 - 139	9.	5.	4.	0.56	0.44
140 - 149	6.	6.	0	1.00	0.
150 - 159	7.	6.	1.	0.86	0.14
160 - 169	9.	9.	0.	1.00	0.
170 - 179	3.	3.	0.	1.00	0.
180 - 189	5.	5.	0.	1.00	0.
190 - 199	1.	1.	0.	1.00	0.
200 - 209	0.	0.	0.	0.	0.
210 - 219	1.	1.	0.	1.00	0.
220 - 229	0.	0.	0.	0.	0.
230 - 239	2.	2.	0.	1.00	0.
240 - 249	0.	0.	0.	0.	0.
250 - UP	0.	0.	0.	0.	0.
TOTAL	209.	175.	34.	0.84	0.16

	MEAN	S D.
CORRECT DOCUMENTS	95.45	44.33
INCORRECT DOCUMENTS	79.71	32.19

TABLE 3   EFFECTIVENESS VS WORDS IN INTERSECTION OF DOCUMENTS AND DISCRIMINATING SET
70 DOCUMENTS IN EACH CATEGORY   DISCRIMINATING WORD SET   2   48 DISCRIMINATING WORDS

SAMPLE DOCUMENTS		CATEGORY A			
NUMBER OF WORDS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0	C.	0.	C.	0.	0.
1	0.	0.	0.	0.	0.
2	2.	2.	0.	1.00	0.
3	5.	4.	1.	0.80	0.20
4	7.	5.	2.	0.71	0.29
5	9.	8.	1.	0.89	0.11
6	15.	13.	2.	0.87	0.13
7	8.	8.	0.	1.0C	0.
8	11.	11.	0.	1.00	0.
9	4.	4.	0.	1.00	0.
10	3.	3.	0.	1.00	0.
11	5.	5.	0.	1.00	0.
12	0.	0.	0.	0.	0.
13	0.	0.	0.	0.	0.
14	1.	1.	0.	1.00	0.
15	0.	0.	0.	0.	0.
16	0.	0.	0.	0.	0.
17	0.	0.	0.	0.	0.
18	0.	0.	0.	0.	0.
19	0.	0.	0.	0.	0.
20	0.	0.	0.	0.	0.
TOTAL	70.	64.	6.	0.91	0.09

	MEAN	S.D.
CORRECT DOCUMENTS	6.77	2.45
INCORRECT DOCUMENTS	4.67	1.11

TABLE 3 EFFECTIVENESS VS WORDS IN INTERSECTION OF DOCUMENTS AND DISCRIMINATING SET
70 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 2 48 DISCRIMINATING WORDS

SAMPLE DOCUMENTS		CATEGORY M			
NUMBER OF WORDS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0	1.	1.	0.	1.00	0.
1	0.	0.	0.	0.	0.
2	2.	2.	0.	1.00	0.
3	5.	3.	2.	0.60	0.40
4	6.	5.	1.	0.83	0.17
5	16.	14.	2.	0.88	0.13
6	10.	7.	3.	0.70	0.30
7	8.	8.	0.	1.00	0.
8	11.	6.	5.	0.55	0.45
9	5.	5.	0.	1.00	0.
10	1.	1.	0.	1.00	0.
11	1.	1.	0.	1.00	0.
12	2.	2.	0.	1.00	0.
13	2.	2.	0.	1.00	0.
14	0.	0.	0.	0.	0.
15	0.	0.	0.	0.	0.
16	0.	0.	0.	0.	0.
17	0.	0.	0.	0.	0.
18	0.	0.	0.	0.	0.
19	0.	0.	0.	0.	0.
20	0.	0.	0.	0.	0.
TOTAL	70.	57.	13.	0.81	0.19

	MEAN	S.D.
CORRECT DOCUMENTS	6.40	2.67
INCORRECT DOCUMENTS	6.00	1.84

TABLE 3 EFFECTIVENESS VS WORDS IN INTERSECTION OF DOCUMENTS AND DISCRIMINATING SET
70 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 2 48 DISCRIMINATING WORDS

SAMPLE DOCUMENTS			CATEGORY P		
NUMBER OF WORDS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0	0.	0.	C.	0.	0.
1	2.	1.	1.	0.50	0.50
2	2.	2.	0.	1.00	0.
3	3.	3.	0.	1.00	0.
4	3.	1.	2.	0.33	0.67
5	13.	9.	4.	0.69	0.31
6	8.	6.	2.	0.75	0.25
7	10.	8.	2.	0.80	0.20
8	4.	2.	2.	0.50	0.50
9	5.	5.	0.	1.00	0.
10	5.	4.	1.	0.80	0.20
11	4.	4.	0.	1.00	0.
12	3.	2.	1.	0.67	0.33
13	3.	3.	0.	1.00	0.
14	2.	2.	0.	1.00	0.
15	1.	1.	0.	1.00	0.
16	1.	1.	C.	1.00	0.
17	0.	0.	0.	0.	0.
18	0.	0.	0.	0.	0.
19	0.	0.	0.	0.	0.
20	0.	0.	0.	0.	0.
TOTAL	69.	54.	15.	0.78	0.22

	MEAN	S.D.
CORRECT DOCUMENTS	7.81	3.53
INCORRECT DOCUMENTS	6.20	2.56

TABLE 3   EFFECTIVENESS VS WORDS IN INTERSECTION OF DOCUMENTS AND DISCRIMINATING SET
70 DOCUMENTS IN EACH CATEGORY   DISCRIMINATING WORD SET   2   48 DISCRIMINATING WORDS

SAMPLE DOCUMENTS		CATEGORY TOTAL			
NUMBER OF WORDS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0	1.	1.	0.	1.00	0.
1	2.	1.	1.	0.50	0.50
2	6.	6.	0.	1.00	0.
3	13.	10.	3.	0.77	0.23
4	16.	11.	5.	0.69	0.31
5	38.	31.	7.	0.82	0.18
6	33.	26.	7.	0.79	0.21
7	26.	24.	2.	0.92	0.09
8	26.	19.	7.	0.73	0.27
9	14.	14.	0.	1.00	0.
10	9.	8.	1.	0.89	0.11
11	10.	10.	0.	1.00	0.
12	5.	4.	1.	0.80	0.20
13	5.	5.	0.	1.00	0.
14	3.	3.	0.	1.00	0.
15	1.	1.	0.	1.00	0.
16	1.	1.	0.	1.00	0.
17	0.	0.	0.	0.	0.
18	0.	0.	0.	0.	0.
19	0.	0.	0.	0.	0.
20	0.	0.	0.	0.	0.
TOTAL	209.	175.	34.	0.84	0.16

	MEAN	S.D.
CORRECT DOCUMENTS	6.97	2.95
INCORRECT DOCUMENTS	5.85	2.17

TABLE 4 EFFECTIVENESS VS RADIUS
70 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 2 48 DISCRIMINATING WORDS

SAMPLE DOCUMENTS			CATEGORY A		
RADIUS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0. - 0.49	12.	8.	4.	0.67	0.33
0.5 - 0.99	21.	20.	1.	0.95	0.05
1.0 - 1.99	37.	36.	1.	0.97	0.03
2.0 - 2.99	0.	0.	0.	0.	0.
3.0 - 3.99	0.	0.	0.	0.	0.
4.0 - 4.99	0.	0.	0.	0.	0.
5.0 - 5.99	0.	0.	0.	0.	0.
6.0 - 6.99	0.	0.	0.	0.	0.
7.0 - 7.99	0.	0.	0.	0.	0.
8.0 - 8.99	0.	0.	0.	0.	0.
9.0 - 9.99	0.	0.	0.	0.	0.
10. - UP	0.	0.	0.	0.	0.
TOTAL	70.	64.	6.	0.91	0.09

	MEAN	S.D.
CORRECT DOCUMENTS	1.06	0.47
INCORRECT DOCUMENTS	0.46	0.37

TABLE 4 EFFECTIVENESS VS RADIUS
70 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 2 48 DISCRIMINATING WORDS

SAMPLE DOCUMENTS			CATEGORY M		
RADIUS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0. - 0.49	26.	21.	5.	0.81	0.19
0.5 - 0.99	33.	26.	7.	0.79	0.21
1.0 - 1.99	10.	9.	1.	0.90	0.10
2.0 - 2.99	1.	1.	0.	1.00	0.
3.0 - 3.99	0.	0.	0.	0.	0.
4.0 - 4.99	0.	0.	0.	0.	0.
5.0 - 5.99	0.	0.	0.	0.	0.
6.0 - 6.99	0.	0.	0.	0.	0.
7.0 - 7.99	0.	0.	0.	0.	0.
8.0 - 8.99	0.	0.	0.	0.	0.
9.0 - 9.99	0.	0.	0.	0.	0.
10. - UP	0.	0.	0.	0.	0.
TOTAL	70.	57.	13.	0.81	0.19

	MEAN	S.D.
CORRECT DOCUMENTS	0.69	0.49
INCORRECT DOCUMENTS	0.62	0.27

TABLE 4 EFFECTIVENESS VS RADIUS
70 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 2 48 DISCRIMINATING WORDS

SAMPLE DOCUMENTS			CATEGORY P		
RADIUS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0. - 0.49	19.	10.	9.	0.53	0.47
0.5 - 0.99	24.	19.	5.	0.79	0.21
1.0 - 1.99	23.	22.	1.	0.96	0.04
2.0 - 2.99	3.	3.	0.	1.00	0.
3.0 - 3.99	0.	0.	0.	0.	0.
4.0 - 4.99	0.	0.	0.	0.	0.
5.0 - 5.99	0.	0.	0.	0.	0.
6.0 - 6.99	0.	0.	0.	0.	0.
7.0 - 7.99	0.	0.	0.	0.	0.
8.0 - 8.99	0.	0.	0.	0.	0.
9.0 - 9.99	0.	0.	0.	0.	0.
10. - UP	0.	0.	0.	0.	0.
TOTAL	69.	54.	15.	0.78	0.22

	MEAN	S.D.
CORRECT DOCUMENTS	1.05	0.56
INCORRECT DOCUMENTS	0.52	0.25

TABLE 4 EFFECTIVENESS VS RADIUS
70 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 2 48 DISCRIMINATING WORDS

SAMPLE DOCUMENTS			CATEGORY TOTAL		
RADIUS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0. - 0.49	57.	39.	18.	0.68	0.32
0.5 - 0.99	78.	65.	13.	0.83	0.17
1.0 - 1.99	70.	67.	3.	0.96	0.04
2.0 - 2.99	4.	4.	0.	1.00	0.
3.0 - 3.99	0.	0.	0.	0.	0.
4.0 - 4.99	0.	0.	0.	0.	0.
5.0 - 5.99	0.	0.	0.	0.	0.
6.0 - 6.99	0.	0.	0.	0.	0.
7.0 - 7.99	0.	0.	0.	0.	0.
8.0 - 8.99	0.	0.	0.	0.	0.
9.0 - 9.99	0.	0.	0.	0.	0.
10. - UP	0.	0.	0.	0.	0.
TOTAL	209.	175.	34.	0.84	0.16

	MEAN	S.D.
CORRECT DOCUMENTS	0.94	0.53
INCORRECT DOCUMENTS	0.55	0.29

```
TABLE 5    DOCUMENT CLASSIFICATION SUMMARY
           70 DOCUMENTS IN EACH CATEGORY   DISCRIMINATING WORD SET    2      48 DISCRIMINATING  WORDS

SAMPLE
                  AUTO    CATEGORY
ACTUAL
CATEGORY         A         M         P        TOTAL

A              64.00      3.00      3.00     70.00

M               2.00     57.00     11.00     70.00

P               4.00     11.00     54.00     69.00

TOTAL          70.00     71.00     68.00    209.00

                    PERCENTAGE

A               0.91      0.04      0.04      1.00

M               0.03      0.81      0.16      1.00

P               0.06      0.16      0.78      1.00

                         SWETS MEASURES               RECALL        RELEVANCE       PRECISION
CATEGORY A           PERTINENT    NOT PERTINENT        RATIO          RATIO           RATIO
 RETRIEVED             0.91            0.04             0.91           0.91            1.00


                         SWETS MEASURES               RECALL        RELEVANCE       PRECISION
CATEGORY M           PERTINENT    NOT PERTINENT        RATIO          RATIO           RATIO
 RETRIEVED             0.81            0.10             0.81           0.80            1.01


                         SWETS MEASURES               RECALL        RELEVANCE       PRECISION
CATEGORY P           PERTINENT    NOT PERTINENT        RATIO          RATIO           RATIO
 RETRIEVED             0.78            0.10             0.78           0.79            0.99
```

TABLE 1 EFFECTIVENESS VS NUMBER OF SENTENCES
70 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 2 48 DISCRIMINATING WORDS

TEST DOCUMENTS			CATEGORY A		
NUMBER OF SENTENCES	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
1	4.	4.	0.	1.00	0.
2	39.	31.	8.	0.79	0.21
3	56.	48.	8.	0.86	0.14
4	55.	46.	9.	0.84	0.16
5	61.	48.	13.	0.79	0.21
6	33.	29.	4.	0.88	0.12
7	24.	20.	4.	0.83	0.17
8	10.	10.	0.	1.00	0.
9	4.	4.	0.	1.00	0.
10	3.	1.	2.	0.33	0.67
11	1.	1.	0.	1.00	0.
12	0.	0.	0.	0.	0.
13	1.	1.	0.	1.00	0.
14	0.	0.	0.	0.	0.
15	0.	0.	0.	0.	0.
TOTAL	291.	243.	48.	0.84	0.16

	MEAN	S.D.
CORRECT DOCUMENTS	4.52	1.94
INCORRECT DOCUMENTS	4.44	1.87

TABLE 1 EFFECTIVENESS VS NUMBER OF SENTENCES
70 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 2 48 DISCRIMINATING WORDS

TEST DOCUMENTS			CATEGORY M		
NUMBER OF SENTENCES	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
1	11.	9.	2.	0.82	0.18
2	26.	18.	8.	0.69	0.31
3	50.	39.	11.	0.78	0.22
4	60.	44.	16.	0.73	0.27
5	55.	39.	16.	0.71	0.29
6	43.	34.	9.	0.79	0.21
7	24.	17.	7.	0.71	0.29
8	23.	15.	8.	0.65	0.35
9	12.	9.	3.	0.75	0.25
10	5.	4.	1.	0.80	0.20
11	3.	3.	0.	1.00	0.
12	1.	1.	0.	1.00	0.
13	0.	0.	0.	0.	0.
14	2.	1.	1.	0.50	0.50
15	0.	0.	0.	0.	0.
TOTAL	315.	233.	82.	0.74	0.26

	MEAN	S.D.
CORRECT DOCUMENTS	4.96	2.29
INCORRECT DOCUMENTS	5.04	2.28

TABLE 1 EFFECTIVENESS VS NUMBER OF SENTENCES
70 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 2 48 DISCRIMINATING WORDS

TEST DOCUMENTS			CATEGORY P		
NUMBER OF SENTENCES	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
1	13.	7.	6.	0.54	0.46
2	65.	40.	25.	0.62	0.38
3	139.	63.	76.	0.45	0.55
4	172.	101.	71.	0.59	0.41
5	158.	101.	57.	0.64	0.36
6	134.	90.	44.	0.67	0.33
7	95.	64.	31.	0.67	0.33
8	66.	51.	15.	0.77	0.23
9	48.	37.	11.	0.77	0.23
10	26.	15.	11.	0.58	0.42
11	10.	8.	2.	0.80	0.20
12	8.	5.	3.	0.63	0.38
13	1.	1.	0.	1.00	0.
14	2.	2.	0.	1.00	0.
15	0.	0.	0.	0.	0.
TOTAL	937.	585.	352.	0.62	0.38

	MEAN	S.D.
CORRECT DOCUMENTS	5.56	2.33
INCORRECT DOCUMENTS	4.89	2.17

TABLE 1 EFFECTIVENESS VS NUMBER OF SENTENCES
70 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 2 48 DISCRIMINATING WORDS

TEST DOCUMENTS		CATEGORY TOTAL			
NUMBER OF SENTENCES	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
1	28.	20.	8.	0.71	0.29
2	130.	89.	41.	0.68	0.32
3	245.	150.	95.	0.61	0.39
4	287.	191.	96.	0.67	0.33
5	274.	188.	86.	0.69	0.31
6	210.	153.	57.	0.73	0.27
7	143.	101.	42.	0.71	0.29
8	99.	76.	23.	0.77	0.23
9	64.	50.	14.	0.78	0.22
10	34.	20.	14.	0.59	0.41
11	14.	12.	2.	0.86	0.14
12	9.	6.	3.	0.67	0.33
13	2.	2.	0.	1.00	0.
14	4.	3.	1.	0.75	0.25
15	0.	0.	0.	0.	0.
TOTAL	1543.	1061.	482.	0.69	0.31

	MEAN	S.D.
CORRECT DOCUMENTS	5.19	2.28
INCORRECT DOCUMENTS	4.87	2.17

TABLE 2 EFFECTIVENESS VS DOCUMENT LENGTH
70 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 2 48 DISCRIMINATING WORDS

TEST DOCUMENTS		CATEGORY A			
NUMBER OF TOKENS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0 - 9	0.	0.	0.	0.	0.
10 - 19	0.	0.	0.	0.	0.
20 - 29	4.	4.	0.	1.00	0.
30 - 39	15.	14.	1.	0.93	0.07
40 - 49	27.	20.	7.	0.74	0.26
50 - 59	32.	27.	5.	0.84	0.16
60 - 69	39.	29.	10.	0.74	0.26
70 - 79	36.	34.	2.	0.94	0.06
80 - 89	32.	24.	8.	0.75	0.25
90 - 99	30.	26.	4.	0.87	0.13
100 - 109	25.	18.	7.	0.72	0.28
110 - 119	14.	13.	1.	0.93	0.07
120 - 129	19.	17.	2.	0.89	0.11
130 - 139	5.	5.	0.	1.00	0.
140 - 149	5.	5.	0.	1.00	0.
150 - 159	2.	2.	0.	1.00	0.
160 - 169	4.	4.	0.	1.00	0.
170 - 179	1.	0.	1.	0.	1.00
180 - 189	1.	1.	0.	1.00	0.
190 - 199	0.	0.	0.	0.	0.
200 - 209	0.	0.	0.	0.	0.
210 - 219	0.	0.	0.	0.	0.
220 - 229	0.	0.	0.	0.	0.
230 - 239	0.	0.	0.	0.	0.
240 - 249	0.	0.	0.	0.	0.
250 - UP	0.	0.	0.	0.	0.
TOTAL	291.	243.	48.	0.84	0.16

	MEAN	S.D.
CORRECT DOCUMENTS	81.63	31.55
INCORRECT DOCUMENTS	78.48	27.47

TABLE 2 EFFECTIVENESS VS DOCUMENT LENGTH
70 DOCUMENTS IN EACH CATEGORY. DISCRIMINATING WORD SET 2 48 DISCRIMINATING WORDS

TEST DOCUMENTS		CATEGORY M			
NUMBER OF TOKENS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0 - 9	0.	0.	0.	0.	0.
10 - 19	2.	2.	0.	1.00	0.
20 - 29	7.	5.	2.	0.71	0.29
30 - 39	14.	11.	3.	0.79	0.21
40 - 49	23.	17.	6.	0.74	0.26
50 - 59	25.	16.	9.	0.64	0.36
60 - 69	37.	29.	8.	0.78	0.22
70 - 79	29.	24.	5.	0.83	0.17
80 - 89	32.	24.	8.	0.75	0.25
90 - 99	24.	17.	7.	0.71	0.29
100 - 109	24.	19.	5.	0.79	0.21
110 - 119	20.	13.	7.	0.65	0.35
120 - 129	24.	18.	6.	0.75	0.25
130 - 139	13.	11.	2.	0.85	0.15
140 - 149	10.	5.	5.	0.50	0.50
150 - 159	9.	8.	1.	0.89	0.11
160 - 169	3.	3.	0.	1.00	0.
170 - 179	7.	3.	4.	0.43	0.57
180 - 189	5.	4.	1.	0.80	0.20
190 - 199	1.	1.	0.	1.00	0.
200 - 209	4.	1.	3.	0.25	0.75
210 - 219	0.	0.	0.	0.	0.
220 - 229	1.	1.	0.	1.00	0.
230 - 239	0.	0.	0.	0.	0.
240 - 249	1.	1.	0.	1.00	0.
250 - UP	0.	0.	0.	0.	0.
TOTAL	315.	233.	82.	0.74	0.26

	MEAN	S.D.
CORRECT DOCUMENTS	91.00	40.84
INCORRECT DOCUMENTS	95.91	43.43

TABLE 2 EFFECTIVENESS VS DOCUMENT LENGTH
70 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 2 48 DISCRIMINATING WORDS

TEST DOCUMENTS		CATEGORY P			
NUMBER OF TOKENS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0 - 9	0.	0.	0.	0.	0.
10 - 19	0.	0.	0.	0.	0.
20 - 29	9.	4.	5.	0.44	0.56
30 - 39	34.	16.	18.	0.47	0.53
40 - 49	44.	19.	25.	0.43	0.57
50 - 59	68.	38.	30.	0.56	0.44
60 - 69	96.	54.	42.	0.56	0.44
70 - 79	86.	47.	39.	0.55	0.45
80 - 89	77.	41.	36.	0.53	0.47
90 - 99	69.	43.	26.	0.62	0.38
100 - 109	63.	43.	20.	0.68	0.32
110 - 119	62.	40.	22.	0.65	C.35
120 - 129	75.	51.	24.	0.68	0.32
130 - 139	56.	37.	19.	0.66	0.34
140 - 149	44.	29.	15.	0.66	0.34
150 - 159	33.	26.	7.	0.79	0.21
160 - 169	25.	21.	4.	0.84	0.16
170 - 179	31.	25.	6.	0.81	0.19
180 - 189	20.	17.	3.	0.85	0.15
190 - 199	16.	15.	1.	0.94	0.06
200 - 209	11.	5.	6.	0.45	0.55
210 - 219	4.	3.	1.	0.75	0.25
220 - 229	4.	3.	1.	0.75	0.25
230 - 239	7.	5.	2.	0.71	0.29
240 - 249	1.	1.	0.	1.00	0.
250 - UP	2.	2.	0.	1.00	0.
TOTAL	937.	585.	352.	0.62	0.38

	MEAN	S.D.
CORRECT DOCUMENTS	110.87	46.64
INCORRECT DOCUMENTS	92.60	41.83

TABLE 2   EFFECTIVENESS VS DOCUMENT LENGTH
70 DOCUMENTS IN EACH CATEGORY   DISCRIMINATING WORD SET   2   48 DISCRIMINATING WORDS

TEST DOCUMENTS NUMBER OF TOKENS	 NUMBER OF DOCUMENTS	 NUMBER OF CORRECT DOCUMENTS	CATEGORY TOTAL NUMBER OF INCORRECT DOCUMENTS	 PERCENTAGE OF CORRECT DOCUMENTS	 PERCENTAGE OF INCCRRECT DOCUMENTS
0 - 9	0.	0.	0.	0.	0.
10 - 19	2.	2.	0.	1.00	0.
20 - 29	20.	13.	7.	0.65	0.35
30 - 39	63.	41.	22.	0.65	0.35
40 - 49	94.	56.	38.	0.60	0.40
50 - 59	125.	81.	44.	0.65	0.35
60 - 69	172.	112.	60.	0.65	0.35
70 - 79	151.	105.	46.	0.70	0.30
80 - 89	141.	89.	52.	0.63	0.37
90 - 99	123.	86.	37.	0.70	0.30
100 - 109	112.	80.	32.	0.71	0.29
110 - 119	96.	66.	30.	0.69	0.31
120 - 129	118.	86.	32.	0.73	0.27
130 - 139	74.	53.	21.	0.72	0.28
140 - 149	59.	39.	20.	0.66	0.34
150 - 159	44.	36.	8.	0.82	0.18
160 - 169	32.	28.	4.	0.88	0.13
170 - 179	39.	28.	11.	0.72	0.28
180 - 189	26.	22.	4.	0.85	0.15
190 - 199	17.	16.	1.	0.94	0.06
200 - 209	15.	6.	9.	0.40	0.60
210 - 219	4.	3.	1.	0.75	0.25
220 - 229	5.	4.	1.	0.80	0.20
230 - 239	7.	5.	2.	0.71	0.29
240 - 249	2.	2.	0.	1.00	0.
250 - UP	2.	2.	0.	1.00	0.
TOTAL	1543.	1061.	482.	0.69	0.31

	MEAN	S.D.
CURRECT DOCUMENTS	99.81	44.20
INCORRECT DOCUMENTS	91.76	41.17

TABLE 3 EFFECTIVENESS VS WORDS IN INTERSECTION OF DOCUMENTS AND DISCRIMINATING SET
70 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 2 48 DISCRIMINATING WORDS

TEST DOCUMENTS			CATEGORY A		
NUMBER OF WORDS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0	1.	0.	1.	0.	1.00
1	1.	0.	1.	0.	1.00
2	7.	3.	4.	0.43	0.57
3	21.	17.	4.	0.81	0.19
4	35.	28.	7.	0.80	0.20
5	40.	30.	10.	0.75	0.25
6	42.	36.	6.	0.86	0.14
7	38.	33.	5.	0.87	0.13
8	36.	32.	4.	0.89	0.11
9	28.	25.	3.	0.89	0.11
10	19.	17.	2.	0.89	0.11
11	11.	11.	0.	1.00	0.
12	7.	7.	0.	1.00	0.
13	2.	2.	0.	1.00	0.
14	3.	2.	1.	0.67	0.33
15	0.	0.	0.	0.	0.
16	0.	0.	0.	0.	0.
17	0.	0.	0.	0.	0.
18	0.	0.	0.	0.	0.
19	0.	0.	0.	0.	0.
20	0.	0.	0.	0.	0.
TOTAL	291.	243.	48.	0.84	0.16

	MEAN	S.D.
CORRECT DOCUMENTS	6.90	2.53
INCORRECT DOCUMENTS	5.48	2.62

TABLE 3 EFFECTIVENESS VS WORDS IN INTERSECTION OF DOCUMENTS AND DISCRIMINATING SET
70 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 2 48 DISCRIMINATING WORDS

TEST DOCUMENTS			CATEGORY M		
NUMBER OF WORDS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0	5.	5.	0.	1.00	0.
1	6.	4.	2.	0.67	0.33
2	19.	11.	8.	0.58	0.42
3	26.	18.	8.	0.69	0.31
4	29.	18.	11.	0.62	0.38
5	34.	24.	10.	0.71	0.29
6	47.	41.	6.	0.87	0.13
7	45.	33.	12.	0.73	0.27
8	41.	31.	10.	0.76	0.24
9	27.	20.	7.	0.74	0.26
10	12.	8.	4.	0.67	0.33
11	11.	9.	2.	0.82	0.18
12	7.	6.	1.	0.86	0.14
13	3.	2.	1.	0.67	0.33
14	0.	0.	0.	0.	0.
15	0.	0.	0.	0.	0.
16	2.	2.	0.	1.00	0.
17	0.	0.	0.	0.	0.
18	0.	0.	0.	0.	0.
19	1.	1.	0.	1.00	0.
20	0.	0.	0.	0.	0.
TOTAL	315.	233.	82.	0.74	0.26

	MEAN	S.D.
CORRECT DOCUMENTS	6.46	2.96
INCORRECT DOCUMENTS	5.93	2.74

TABLE 3 EFFECTIVENESS VS WORDS IN INTERSECTION OF DOCUMENTS AND DISCRIMINATING SET
70 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 2 48 DISCRIMINATING WORDS

	TEST DOCUMENTS		CATEGORY P		
NUMBER OF WORDS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0	2.	0.	2.	0.	1,00
1	9.	1.	8.	0.11	0.89
2	20.	9.	11.	0.45	0.55
3	38.	12.	26.	0.32	0.68
4	78.	40.	38.	0.51	0.49
5	97.	44.	53.	0.45	0.55
6	112.	64.	48.	0.57	0.43
7	132.	86.	46.	0.65	0.35
8	102.	73.	29.	0.72	0.28
9	103.	67.	36.	0.65	0.35
10	75.	61.	14.	0.81	0.19
11	48.	34.	14.	0.71	0.29
12	49.	37.	12.	0.76	0.24
13	33.	23.	10.	0.70	0.30
14	18.	14.	4.	0.78	0.22
15	9.	9.	0.	1.00	0.
16	4.	3.	1.	0.75	0.25
17	5.	5.	C.	1.00	0.
18	2.	2.	0.	1.0C	0.
19	1.	1.	0.	1.00	0.
20	0.	0.	C.	0.	0.
TOTAL	937.	585.	352.	0.62	0.38

	MEAN	S.D.
CORRECT DOCUMENTS	8.30	3.11
INCORRECT DOCUMENTS	6:62	2.96

TABLE 3 EFFECTIVENESS VS WORDS IN INTERSECTION OF DOCUMENTS AND DISCRIMINATING SET
70 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 2 48 DISCRIMINATING WORDS

TEST DOCUMENTS			CATEGORY TOTAL		
NUMBER OF WORDS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0	8.	5.	3.	0.63	0.38
1	16.	5.	11.	0.31	0.69
2	46.	23.	23.	0.50	0.50
3	85.	47.	38.	0.55	0.45
4	142.	86.	56.	0.61	0.39
5	171.	98.	73.	0.57	0.43
6	201.	141.	60.	0.70	0.30
7	215.	152.	63.	0.71	0.29
8	179.	136.	43.	0.76	0.24
9	158.	112.	46.	0.71	0.29
10	106.	86.	20.	0.81	0.19
11	70.	54.	16.	0.77	0.23
12	63.	50.	13.	0.79	0.21
13	38.	27.	11.	0.71	0.29
14	21.	16.	5.	0.76	0.24
15	9.	9.	0.	1.00	0.
16	6.	5.	1.	0.83	0.17
17	5.	5.	0.	1.00	0.
18	2.	2.	0.	1.00	0.
19	2.	2.	0.	1.00	0.
20	0.	0.	0.	0.	0.
TOTAL	1543.	1061.	482.	0.69	0.31

	MEAN	S.D.
CORRECT DOCUMENTS	7.57	3.06
INCORRECT DOCUMENTS	6.39	2.92

TABLE 4   EFFECTIVENESS VS RADIUS
70 DOCUMENTS IN EACH CATEGORY   DISCRIMINATING WORD SET   2   48 DISCRIMINATING WORDS

	TEST DOCUMENTS		CATEGORY A		
RADIUS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0. - 0.49	66.	32.	34.	0.48	0.52
0.5 - 0.99	92.	82.	10.	0.89	0.11
1.0 - 1.99	103.	99.	4.	0.96	0.04
2.0 - 2.99	26.	26.	0.	1.00	0.
3.0 - 3.99	4.	4.	0.	1.00	0.
4.0 - 4.99	0.	0.	0.	0.	0.
5.0 - 5.99	0.	0.	0.	0.	0.
6.0 - 6.99	0.	0.	0.	0.	0.
7.0 - 7.99	0.	0.	0.	0.	0.
8.0 - 8.99	0.	0.	0.	0.	0.
9.0 - 9.99	0.	0.	0.	0.	0.
10. - UP	0.	0.	0.	0.	0.
TOTAL	291.	243.	48.	0.84	0.16

	MEAN	S.D.
CORRECT DOCUMENTS	1.17	0.65
INCORRECT DOCUMENTS	0.45	0.41

TABLE 4 EFFECTIVENESS VS RADIUS
70 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 2 48 DISCRIMINATING WORDS

	TEST DOCUMENTS		CATEGORY M		
RADIUS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0. - 0.49	138.	100.	38.	0.72	0.28
0.5 - 0.99	117.	84.	33.	0.72	0.28
1.0 - 1.99	53.	42.	11.	0.79	0.21
2.0 - 2.99	7.	7.	0.	1.00	0.
3.0 - 3.99	0.	0.	0.	0.	0.
4.0 - 4.99	0.	0.	0.	0.	0.
5.0 - 5.99	0.	0.	0.	0.	0.
6.0 - 6.99	0.	0.	0.	0.	0.
7.0 - 7.99	0.	0.	0.	0.	0.
8.0 - 8.99	0.	0.	0.	0.	0.
9.0 - 9.99	0.	0.	0.	0.	0.
10. - UP	0.	0.	0.	0.	0.
TOTAL	315.	233.	82.	0.74	0.26

	MEAN	S.D.
CORRECT DOCUMENTS	0.70	0.52
INCORRECT DOCUMENTS	0.63	0.40

TABLE 4 EFFECTIVENESS VS RADIUS
70 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 2 48 DISCRIMINATING WORDS

TEST DOCUMENTS RADIUS	 NUMBER OF DOCUMENTS	 NUMBER OF CORRECT DOCUMENTS	CATEGORY P NUMBER OF INCORRECT DOCUMENTS	 PERCENTAGE OF CORRECT DOCUMENTS	 PERCENTAGE OF INCORRECT DOCUMENTS
0. - 0.49	257.	78.	179.	0.30	0.70
0.5 - 0.99	347.	212.	135.	0.61	0.39
1.0 - 1.99	250.	215.	35.	0.86	0.14
2.0 - 2.99	66.	63.	3.	0.95	0.05
3.0 - 3.99	10.	10.	0.	1.00	0.
4.0 - 4.99	5.	5.	0.	1.00	0.
5.0 - 5.99	2.	2.	0.	1.00	0.
6.0 - 6.99	0.	0.	0.	0.	0.
7.0 - 7.99	0.	0.	0.	0.	0.
8.0 - 8.99	0.	0.	0.	0.	0.
9.0 - 9.99	0.	0.	0.	0.	0.
10. - UP	0.	0.	0.	0.	0.
TOTAL	937.	585.	352.	0.62	0.38

	MEAN	S.D.
CORRECT DOCUMENTS	1.20	0.79
INCORRECT DOCUMENTS	0.56	0.38

TABLE 4 EFFECTIVENESS VS RADIUS
70 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 2 48 DISCRIMINATING WORDS

	TEST DOCUMENTS		CATEGORY TOTAL		
RADIUS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0. - 0.49	461.	210.	251.	0.46	0.54
0.5 - 0.99	556.	378.	178.	0.68	0.32
1.0 - 1.99	406.	356.	50.	0.88	0.12
2.0 - 2.99	99.	96.	3.	0.97	0.03
3.0 - 3.99	14.	14.	0.	1.00	0.
4.0 - 4.99	5.	5.	0.	1.00	0.
5.0 - 5.99	2.	2.	0.	1.00	0.
6.0 - 6.99	0.	0.	0.	0.	0.
7.0 - 7.99	0.	0.	0.	0.	0.
8.0 - 8.99	0.	0.	0.	0.	0.
9.0 - 9.99	0.	0.	0.	0.	0.
10. - UP	0.	0.	0.	0.	0.
TOTAL	1543.	1061.	482.	0.69	0.31

	MEAN	S.D.
CORRECT DOCUMENTS	1.08	0.73
INCORRFCT DOCUMENTS	0.56	0.39

```
TABLE 5   DOCUMENT CLASSIFICATION SUMMARY
          70 DOCUMENTS IN EACH CATEGORY   DISCRIMINATING WORD SET    2      48 DISCRIMINATING  WORDS

  TEST
               AUTO   CATEGORY
ACTUAL
CATEGORY      A        M        P       TOTAL

A           243.00    28.00    20.00   291.00

M            29.00   233.00    53.00   315.00

P            71.00   281.00   585.00   937.00

TOTAL       343.00   542.00   658.00  1543.00

                  PERCENTAGE

A             0.84     0.10     0.07     1.00

M             0.09     0.74     0.17     1.00

P             0.08     0.30     0.62     1.00

                      SWETS MEASURES              RECALL         RELEVANCE        PRECISION
CATEGORY A        PERTINENT   NOT PERTINENT       RATIO           RATIO            RATIO
 RETRIEVED          0.84          0.07             0.84            0.71             1.18


                      SWETS MEASURES              RECALL         RELEVANCE        PRECISION
CATEGORY M        PERTINENT   NOT PERTINENT       RATIO           RATIO            RATIO
 RETRIEVED          0.74          0.21             0.74            0.43             1.72


                      SWETS MEASURES              RECALL         RELEVANCE        PRECISION
CATEGORY P        PERTINENT   NOT PERTINENT       RATIO           RATIO            RATIO
 RETRIEVED          0.62          0.09             0.62            0.89             0.70
```

TABLE 1 EFFECTIVENESS VS NUMBER OF SENTENCES
140 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 2 48 DISCRIMINATING WORDS

SAMPLE DOCUMENTS			CATEGORY A		
NUMBER OF SENTENCES	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
1	3.	3.	0.	1.00	0.
2	20.	16.	4.	0.80	0.20
3	32.	29.	3.	0.91	0.09
4	18.	17.	1.	0.94	0.06
5	29.	23.	6.	0.79	0.21
6	15.	13.	2.	0.87	0.13
7	12.	12.	0.	1.00	0.
8	6.	6.	0.	1.00	0.
9	2.	2.	0.	1.00	0.
10	1.	0.	1.	0.	1.00
11	1.	1.	0.	1.00	0.
12	0.	0.	0.	0.	0.
13	1.	1.	0.	1.00	0.
14	0.	0.	0.	0.	0.
15	0.	0.	0.	0.	0.
TOTAL	140.	123.	17.	0.88	0.12

	MEAN	S.D.
CORRECT DOCUMENTS	4.53	2.10
INCORRECT DOCUMENTS	4.29	1.99

TABLE 1 EFFECTIVENESS VS NUMBER OF SENTENCES
140 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 2 48 DISCRIMINATING WORDS

SAMPLE DOCUMENTS		CATEGORY M			
NUMBER OF SENTENCES	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
1	3.	3.	0.	1.00	0.
2	7.	5.	2.	0.71	0.29
3	24.	18.	6.	0.75	0.25
4	24.	20.	4.	0.83	0.17
5	19.	16.	3.	0.84	0.16
6	27.	23.	4.	0.85	0.15
7	14.	10.	4.	0.71	0.29
8	13.	12.	1.	0.92	0.08
9	5.	4.	1.	0.80	0.20
10	3.	3.	0.	1.00	0.
11	1.	1.	0.	1.00	0.
12	0.	0.	0.	0.	0.
13	0.	0.	0.	0.	0.
14	0.	0.	0.	0.	0.
15	0.	0.	0.	0.	0.
TOTAL	140.	115.	25.	0.82	0.18

	MEAN	S.D.
CORRECT DOCUMENTS	5.29	2.13
INCORRECT DOCUMENTS	4.88	1.90

TABLE 1 EFFECTIVENESS VS NUMBER OF SENTENCES
140 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 2 48 DISCRIMINATING WORDS

SAMPLE DOCUMENTS		CATEGORY P			
NUMBER OF SENTENCES	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
1	1.	1.	0.	1.00	0.
2	6.	4.	2.	0.67	0.33
3	26.	20.	6.	0.77	0.23
4	25.	20.	5.	0.80	0.20
5	25.	22.	3.	0.88	0.12
6	19.	17.	2.	0.89	0.11
7	20.	17.	3.	0.85	0.15
8	6.	6.	0.	1.00	0.
9	5.	5.	0.	1.00	0.
10	3.	3.	0.	1.00	0.
11	3.	3.	0.	1.00	0.
12	1.	1.	0.	1.00	0.
13	0.	0.	0.	0.	0.
14	0.	0.	0.	0.	0.
15	0.	0.	0.	0.	0.
TOTAL	140.	119.	21.	0.85	0.15

	MEAN	S.D.
CORRECT DOCUMENTS	5.45	2.21
INCORRECT DOCUMENTS	4.29	1.55

TABLE 1 EFFECTIVENESS VS NUMBER OF SENTENCES
140 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 2 48 DISCRIMINATING WORDS

SAMPLE DOCUMENTS			CATEGORY TOTAL		
NUMBER OF SENTENCES	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
1	7.	7.	0.	1.00	0.
2	33.	25.	8.	0.76	0.24
3	82.	67.	15.	0.82	0.18
4	67.	57.	10.	0.85	0.15
5	73.	61.	12.	0.84	0.16
6	61.	53.	8.	0.87	0.13
7	46.	39.	7.	0.85	0.15
8	25.	24.	1.	0.96	0.04
9	12.	11.	1.	0.92	0.08
10	7.	6.	1.	0.86	0.14
11	5.	5.	0.	1.00	0.
12	1.	1.	0.	1.00	0.
13	1.	1.	0.	1.00	0.
14	0.	0.	0.	0.	0.
15	0.	0.	0.	0.	0.
TOTAL	420.	357.	63.	0.85	0.15

	MEAN	S.D.
CORRECT DOCUMENTS	5.08	2.18
INCORRECT DOCUMENTS	4.52	1.84

TABLE 2 EFFECTIVENESS VS DOCUMENT LENGTH
140 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 2 48 DISCRIMINATING WORDS

SAMPLE DOCUMENTS NUMBER OF TOKENS	 NUMBER OF DOCUMENTS	 NUMBER OF CORRECT DOCUMENTS	CATEGORY A NUMBER OF INCORRECT DOCUMENTS	 PERCENTAGE OF CORRECT DOCUMENTS	 PERCENTAGE OF INCORRECT DOCUMENTS
0 - 9	0.	0.	0.	0.	0.
10 - 19	0.	0.	0.	0.	0.
20 - 29	1.	1.	0.	1.00	0.
30 - 39	6.	6.	0.	1.00	0.
40 - 49	20.	15.	5.	0.75	0.25
50 - 59	14.	11.	3.	0.79	0.21
60 - 69	15.	14.	1.	0.93	0.07
70 - 79	20.	20.	0.	1.00	0.
80 - 89	11.	10.	1.	0.91	0.09
90 - 99	13.	11.	2.	0.85	0.15
100 - 109	12.	9.	3.	0.75	0.25
110 - 119	5.	5.	0.	1.00	0.
120 - 129	10.	10.	0.	1.00	0.
130 - 139	4.	3.	1.	0.75	0.25
140 - 149	1.	1.	0.	1.00	0.
150 - 159	3.	3.	0.	1.00	0.
160 - 169	2.	2.	0.	1.00	0.
170 - 179	1.	0.	1.	0.	1.00
180 - 189	2.	2.	0.	1.00	0.
190 - 199	0.	0.	0.	0.	0.
200 - 209	0.	0.	0.	0.	0.
210 - 219	0.	0.	0.	0.	0.
220 - 229	0.	0.	0.	0.	0.
230 - 239	0.	0.	0.	0.	0.
240 - 249	0.	0.	0.	0.	0.
250 - UP	0.	0.	0.	0.	0.
TOTAL	140.	123,	17.	0.88	0.12

	MEAN	S.D.
CORRECT DOCUMENTS	83.16	33.91
INCORRECT DOCUMENTS	79.65	36.27

TABLE 2 EFFECTIVENESS VS DOCUMENT LENGTH
140 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 2 48 DISCRIMINATING WORDS

SAMPLE DOCUMENTS			CATEGORY M		
NUMBER OF TOKENS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	FERCENTAGE OF INCORRECT DOCUMENTS
0 - 9	0.	0.	0.	0.	0.
10 - 19	1.	1.	0.	1.00	0.
20 - 29	1.	1.	0.	1.00	0.
30 - 39	3.	2.	1.	0.67	0.33
40 - 49	8.	6.	2.	0.75	0.25
50 - 59	14.	10.	4.	0.71	0.29
60 - 69	11.	11.	0.	1.00	0.
70 - 79	13.	11.	2.	0.85	0.15
80 - 89	17.	13.	4.	0.76	0.24
90 - 99	12.	9.	3.	0.75	0.25
100 - 109	9.	9.	0.	1.00	0.
110 - 119	13.	10.	3.	0.77	0.23
120 - 129	10.	10.	0.	1.00	0.
130 - 139	10.	6.	4.	0.60	0.40
140 - 149	5.	4.	1.	0.80	0.20
150 - 159	6.	6.	0.	1.00	0.
160 - 169	1.	1.	0.	1.00	0.
170 - 179	4.	3.	1.	0.75	0.25
180 - 189	2.	2.	0.	1.00	0.
190 - 199	0.	0.	0.	0.	0.
200 - 209	0.	0.	0.	0.	0.
210 - 219	0.	0.	0.	0.	0.
220 - 229	0.	0.	0.	0.	0.
230 - 239	0.	0.	0.	0.	0.
240 - 249	0.	0.	0.	0.	0.
250 - UP	0.	0.	0.	0.	0.
TOTAL	140.	115.	25.	0.82	0.18

	MEAN	S.D.
CORRECT DOCUMENTS	95.97	37.33
INCORRECT DOCUMENTS	93.20	35.85

TABLE 2 EFFECTIVENESS VS DOCUMENT LENGTH
140 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 2 48 DISCRIMINATING WORDS

SAMPLE DOCUMENTS			CATEGORY P		
NUMBER OF TOKENS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0 - 9	0.	0.	0.	0.	0.
10 - 19	0.	0.	0.	0.	0.
20 - 29	1.	1.	0.	1.00	0.
30 - 39	4.	3.	1.	0.75	0.25
40 - 49	7.	6.	1.	0.86	0.14
50 - 59	11.	6.	5.	0.55	0.45
60 - 69	21.	18.	3.	0.86	0.14
70 - 79	13.	11.	2.	0.85	0.15
80 - 89	13.	10.	3.	0.77	0.23
90 - 99	5.	5.	0.	1.00	0.
100 - 109	10.	9.	1.	0.90	0.10
110 - 119	6.	5.	1.	0.83	0.17
120 - 129	7.	6.	1.	0.86	0.14
130 - 139	14.	12.	2.	0.86	0.14
140 - 149	4.	4.	0.	1.00	0.
150 - 159	7.	6.	1.	0.86	0.14
160 - 169	3.	3.	0.	1.00	0.
170 - 179	5.	5.	0.	1.00	0.
180 - 189	3.	3.	0.	1.90	0.
190 - 199	2.	2.	0.	1.00	0.
200 - 209	1.	1.	0.	1.00	0.
210 - 219	0.	0.	0.	0.	0.
220 - 229	0.	0.	0.	0.	0.
230 - 239	3.	3.	0.	1.00	0.
240 - 249	0.	0.	0.	0.	0.
250 - UP	0.	0.	0.	0.	0.
TOTAL	140.	119.	21.	0.85	0.15

	MEAN	S.D.
CORRECT DOCUMENTS	105.84	47.94
INCORRECT DOCUMENTS	81.95	32.93

TABLE 2 EFFECTIVENESS VS DOCUMENT LENGTH
140 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 2 48 DISCRIMINATING WORDS

SAMPLE DOCUMENTS			CATEGORY TOTAL		
NUMBER OF TOKENS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0 - 9	0.	0.	0.	0.	0.
10 - 19	1.	1.	0.	1.00	0.
20 - 29	3.	3.	0.	1.00	0.
30 - 39	13.	11.	2.	0.85	0.15
40 - 49	35.	27.	8.	0.77	0.23
50 - 59	39.	27.	12.	0.69	0.31
60 - 69	47.	43.	4.	0.91	0.09
70 - 79	46.	42.	4.	0.91	0.09
80 - 89	41.	33.	8.	0.80	0.20
90 - 99	30.	25.	5.	0.83	0.17
100 - 109	31.	27.	4.	0.87	0.13
110 - 119	24.	20.	4.	0.83	0.17
120 - 129	27.	26.	1.	0.96	0.04
130 - 139	28.	21.	7.	0.75	0.25
140 - 149	10.	9.	1.	0.90	0.10
150 - 159	16.	15.	1.	0.94	0.06
160 - 169	6.	6.	0.	1.00	0.
170 - 179	10.	8.	2.	0.80	0.20
180 - 189	7.	7.	0.	1.00	0.
190 - 199	2.	2.	0.	1.00	0.
200 - 209	1.	1.	0.	1.00	0.
210 - 219	0.	0.	0.	0.	0.
220 - 229	0.	0.	0.	0.	0.
230 - 239	3.	3.	0.	1.00	0.
240 - 249	0.	0.	0.	0.	0.
250 - UP	0.	0.	0.	0.	0.
TOTA	420.	357.	63.	0.85	0.15

	MEAN	S.D.
CORRECT DOCUMENTS	94.85	41.21
INCORRECT DOCUMENTS	85.79	35.55

TABLE 3 EFFECTIVENESS VS WORDS IN INTERSECTION OF DOCUMENTS AND DISCRIMINATING SET
140 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 2 48 DISCRIMINATING WORDS

SAMPLE DOCUMENTS		CATEGORY A			
NUMBER OF WORDS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0	0.	0.	0.	0.	0.
1	0.	0.	0.	0.	0.
2	5.	3.	2.	0.60	0.40
3	10.	9.	1.	0.90	0.10
4	13.	11.	2.	0.85	0.15
5	22.	19.	3.	0.86	0.14
6	22.	18.	4.	0.82	0.18
7	12.	11.	1.	0.92	0.08
8	19.	17.	2.	0.89	0.11
9	9.	9.	0.	1.00	0.
10	12.	11.	1.	0.92	0.08
11	8.	8.	0.	1.00	0.
12	4.	4.	0.	1.00	0.
13	1.	1.	0.	1.00	0.
14	3.	2.	1.	0.67	0.33
15	0.	0.	0.	0.	0.
16	0.	0.	0.	0.	0.
17	0.	0.	0.	0.	0.
18	0.	0.	0.	0.	0.
19	0.	0.	0.	0.	0.
20	0.	0.	0.	0.	0.
TOTAL	140.	123.	17.	0.88	0.12

	MEAN	S.D.
CORRECT DOCUMENTS	7.00	2.74
INCORRECT DOCUMENTS	5.94	2.88

TABLE 3 EFFECTIVENESS VS WORDS IN INTERSECTION OF DOCUMENTS AND DISCRIMINATING SET
140 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 2 48 DISCRIMINATING WORDS

SAMPLE DOCUMENTS			CATEGORY M		
NUMBER OF WORDS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PEPCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0	0.	0.	0.	0.	0.
1	3.	2.	1.	0.67	0.33
2	7.	3.	4.	0.43	0.57
3	5.	5.	0.	1.00	0.
4	15.	11.	4.	0.73	0.27
5	25.	20.	5.	0.80	0.20
6	20.	17.	3.	0.85	0.15
7	14.	11.	3.	0.79	0.21
8	22.	19.	3.	0.86	0.14
9	10.	9.	1.	0.90	0.10
10	6.	5.	1.	0.83	0.17
11	5.	5.	0.	1.00	0.
12	4.	4.	0.	1.00	0.
13	3.	3.	0.	1.00	0.
14	0.	0.	0.	0.	0.
15	0.	0.	0.	0.	0.
16	1.	1.	0.	1.00	0.
17	0.	0.	0.	0.	0.
18	0.	0.	0.	0.	0.
19	0.	0.	0.	0.	0.
20	0.	0.	0.	0.	0.
TOTAL	140.	115.	25.	0.82	0.18

	MEAN	S.D.
CORRECT DOCUMENTS	6.84	2.78
INCORRECT DOCUMENTS	5.28	2.34

TABLE 3 EFFECTIVENESS VS WORDS IN INTERSECTION OF DOCUMENTS AND DISCRIMINATING SET
140 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 2 48 DISCRIMINATING WORDS

SAMPLE DOCUMENTS		CATEGORY P			
NUMBER OF WORDS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0	0.	0.	0.	0.	0.
1	0.	0.	0.	0.	0.
2	3.	1.	2.	0.33	0.67
3	7.	6.	1.	0.86	0.14
4	15.	10.	5.	0.67	0.33
5	12.	11.	1.	0.92	0.08
6	19.	16.	3.	0.84	0.16
7	16.	15.	1.	0.94	0.06
8	14.	13.	1.	0.93	0.07
9	16.	13.	3.	0.81	0.19
10	9.	8.	1.	0.89	0.11
11	8.	7.	1.	0.88	0.13
12	6.	5.	1.	0.83	0.17
13	6.	5.	1.	0.83	0.17
14	5.	5.	0.	1.00	0.
15	2.	2.	0.	1.00	0.
16	0.	0.	0.	0.	0.
17	2.	2.	0.	1.00	0.
18	0.	0.	0.	0.	0.
19	0.	0.	0.	0.	0.
20	0.	0.	0.	0.	0.
TOTAL	140.	119.	21.	0.85	0.15

	MEAN	S.D.
CORRECT DOCUMENTS	8.01	3.28
INCORRECT DOCUMENTS	6.57	3.19

TABLE 3 EFFECTIVENESS VS WORDS IN INTERSECTION OF DOCUMENTS AND DISCRIMINATING SET
140 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 2 48 DISCRIMINATING WORDS

SAMPLE DOCUMENTS			CATEGORY TOTAL		
NUMBER OF WORDS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0	0.	0.	0.	0.	0.
1	3.	2.	1.	0.67	0.33
2	15.	7.	8.	0.47	0.53
3	22.	20.	2.	0.91	0.09
4	43.	32.	11.	0.74	0.26
5	59.	50.	9.	0.85	0.15
6	61.	51.	10.	0.84	0.16
7	42.	37.	5.	0.88	0.12
8	55.	49.	6.	0.89	0.11
9	35.	31.	4.	0.89	0.11
10	27.	24.	3.	0.89	0.11
11	21.	20.	1.	0.95	0.05
12	14.	13.	1.	0.93	0.07
13	10.	9.	1.	0.90	0.10
14	8.	7.	1.	0.88	0.13
15	2.	2.	0.	1.00	0.
16	1.	1.	0.	1.00	0.
17	2.	2.	0.	1.00	0.
18	0.	0.	0.	0.	0.
19	0.	0.	0.	0.	0.
20	0.	0.	0.	0.	0.
TOTAL	420.	357.	63.	0.85	0.15

	MEAN	S.D.
CORRECT DOCUMENTS	7.29	2.99
INCORRECT DOCUMENTS	5.89	2.85

TABLE 4 EFFECTIVENESS VS RADIUS
140 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 2 48 DISCRIMINATING WORDS

SAMPLE DOCUMENTS			CATEGORY A		
RADIUS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0. - 0.49	26.	17.	9.	0.65	0.35
0.5 - 0.99	30.	28.	2.	0.93	0.07
1.0 - 1.99	64.	58.	6.	0.91	0.09
2.0 - 2.99	19.	19.	0.	1.00	0.
3.0 - 3.99	1.	1.	0.	1.00	0.
4.0 - 4.99	0.	0.	0.	0.	0.
5.0 - 5.99	0.	0.	0.	0.	0.
6.0 - 6.99	0.	0.	0.	0.	0.
7.0 - 7.99	0.	0.	0.	0.	0.
8.0 - 8.99	0.	0.	0.	0.	0.
9.0 - 9.99	0.	0.	0.	0.	0.
10. - UP	0.	0.	0.	0.	0.
TOTAL	140.	123.	17.	0.88	0.12

	MEAN	S.D.
CORRECT DOCUMENTS	1.29	0.67
INCORRECT DOCUMENTS	0.72	0.60

TABLE 4 EFFECTIVENESS VS RADIUS

140 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 2 48 DISCRIMINATING WORDS

SAMPLE DOCUMENTS			CATEGORY M		
RADIUS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0. - 0.49	35.	22.	13.	0.63	0.37
0.5 - 0.99	48.	37.	11.	0.77	0.23
1.0 - 1.99	44.	43.	1.	0.98	0.02
2.0 - 2.99	12.	12.	0.	1.00	0.
3.0 - 3.99	1.	1.	0.	1.00	0.
4.0 - 4.99	0.	0.	0.	0.	0.
5.0 - 5.99	0.	0.	0.	0.	0.
6.0 - 6.99	0.	0.	0.	0.	0.
7.0 - 7.99	0.	0.	0.	0.	0.
8.0 - 8.99	0.	0.	0.	0.	0.
9.0 - 9.99	0.	0.	0.	0.	0.
10. - UP	0.	0.	0.	0.	0.
TOTAL	140.	115.	25.	0.82	0.18

	MEAN	S.D.
CORRECT DOCUMENTS	1.11	0.62
INCORRECT DOCUMENTS	0.52	0.30

TABLE 4 EFFECTIVENESS VS RADIUS

140 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 2 48 DISCRIMINATING WORDS

SAMPLE DOCUMENTS			CATEGORY P		
RADIUS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0. - 0.49	27.	18.	9.	0.67	0.33
0.5 - 0.99	51.	40.	11.	0.78	0.22
1.0 - 1.99	46.	45.	1.	0.98	0.02
2.0 - 2.99	12.	12.	0.	1.00	0.
3.0 - 3.99	4.	4.	0.	1.00	0.
4.0 - 4.99	0.	0.	0.	0.	0.
5.0 - 5.99	0.	0.	0.	0.	0.
6.0 - 6.99	0.	0.	0.	0.	0.
7.0 - 7.99	0.	0.	0.	0.	0.
8.0 - 8.99	0.	0.	0.	0.	0.
9.0 - 9.99	0.	0.	0.	0.	0.
10. - UP	0.	0.	0.	0.	0.
TOTAL	140.	119.	21.	0.85	0.15

	MEAN	S.D.
CORRECT DOCUMENTS	1.20	0.73
INCORRECT DOCUMENTS	0.54	0.26

TABLE 4 EFFECTIVENESS VS RADIUS
140 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 2 48 DISCRIMINATING WORDS

	SAMPLE DOCUMENTS		CATEGORY TOTAL		
RADIUS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0. - 0.49	88.	57.	31.	0.65	0.35
0.5 - 0.99	129.	105.	24.	0.81	0.19
1.0 - 1.99	154.	146.	8.	0.95	0.05
2.0 - 2.99	43.	43.	0.	1.00	0.
3.0 - 3.99	6.	6.	0.	1.00	0.
4.0 - 4.99	0.	0.	0.	0.	0.
5.0 - 5.99	0.	0.	0.	0.	0.
6.0 - 6.99	0.	0.	0.	0.	0.
7.0 - 7.99	0.	0.	0.	0.	0.
8.0 - 8.99	0.	0.	0.	0.	0.
9.0 - 9.99	0.	0.	0.	0.	0.
10. - UP	0.	0.	0.	0.	0.
TOTAL	420.	357.	63.	0.85	0.15

	MEAN	S.D.
CORRECT DOCUMENTS	1.20	0.68
INCORRECT DOCUMENTS	0.58	0.40

TABLE 5 DOCUMENT CLASSIFICATION SUMMARY
140 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 2 48 DISCRIMINATING WORDS

SAMPLE

	AUTO	CATEGORY		
ACTUAL CATEGORY	A	M	P	TOTAL
A	123.00	4.00	13.00	140.00
M	5.00	115.00	20.00	140.00
P	7.00	14.00	119.00	140.00
TOTAL	135.00	133.00	152.00	420.00
	PERCENTAGE			
A	0.88	0.03	0.09	1.00
M	0.04	0.82	0.14	1.00
P	0.05	0.10	0.85	1.00

	SWETS MEASURES		RECALL	RELEVANCE	PRECISION
CATEGORY A RETRIEVED	PERTINENT	NOT PERTINENT	RATIO	RATIO	RATIO
	0.88	0.04	0.88	0.91	0.96

	SWETS MEASURES		RECALL	RELEVANCE	PRECISION
CATEGORY M RETRIEVED	PERTINENT	NOT PERTINENT	RATIO	RATIO	RATIO
	0.82	0.06	0.82	0.86	0.95

	SWETS MEASURES		RECALL	RELEVANCE	PRECISION
CATEGORY P RETRIEVED	PERTINENT	NOT PERTINENT	RATIO	RATIO	RATIO
	0.85	0.12	0.85	0.78	1.09

TABLE 1 EFFECTIVENESS VS NUMBER OF SENTENCES
140 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 2 48 DISCRIMINATING WORDS

TEST DOCUMENTS		CATEGORY A			
NUMBER OF SENTENCES	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
1	2.	2.	0.	1.00	0.
2	24.	20.	4.	0.83	0.17
3	43.	39.	4.	0.91	0.09
4	53.	45.	5.	0.91	0.09
5	42.	38.	4.	0.90	0.10
6	27.	24.	3.	0.89	0.11
7	17.	14.	3.	0.62	0.18
8	9.	7.	2.	0.78	0.22
9	2.	2.	0.	1.00	0.
10	2.	1.	1.	0.50	0.50
11	0.	0.	0.	0.	0.
12	1.	1.	0.	1.00	0.
13	0.	0.	0.	0.	0.
14	0.	0.	0.	0.	0.
15	0.	0.	0.	0.	0.
TOTAL	222.	196.	26.	0.88	0.12

	MEAN	S.D.
CORRECT DOCUMENTS	4.48	1.77
INCORRECT DOCUMENTS	4.81	2.09

TABLE 1 EFFECTIVENESS VS NUMBER OF SENTENCES
140 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 2 48 DISCRIMINATING WORDS

TEST DOCUMENTS		CATEGORY M			
NUMBER OF SENTENCES	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
1	9.	6.	3.	0.67	0.33
2	26.	16.	10.	0.62	0.38
3	35.	24.	11.	0.69	0.31
4	52.	38.	14.	0.73	0.27
5	45.	31.	14.	0.67	0.31
6	25.	19.	6.	0.76	0.24
7	17.	14.	3.	0.82	0.18
8	16.	12.	4.	0.75	0.25
9	13.	11.	2.	0.85	0.15
10	2.	2.	0.	1.00	0.
11	2.	2.	0.	1.00	0.
12	1.	1.	0.	1.00	0.
13	0.	0.	0.	0.	0.
14	2.	2.	0.	1.00	0.
15	0.	0.	0.	0.	0.
TOTAL	245.	178.	67.	0.73	0.27

	MEAN	S.D.
CORRECT DOCUMENTS	5.09	2.45
INCORRECT DOCUMENTS	4.31	1.93

TABLE 1 EFFECTIVENESS VS NUMBER OF SENTENCES
140 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 2 48 DISCRIMINATING WORDS

TEST DOCUMENTS			CATEGORY P		
NUMBER OF SENTENCES	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
1	12.	5.	7.	0.42	0.58
2	65.	43.	22.	0.66	0.34
3	129.	82.	47.	0.64	0.36
4	158.	116.	42.	0.73	0.27
5	145.	112.	33.	0.77	0.23
6	123.	89.	34.	0.72	0.28
7	82.	62.	20.	0.76	0.24
8	63.	51.	12.	0.81	0.19
9	45.	40.	5.	0.89	0.11
10	25.	20.	5.	0.80	0.20
11	9.	8.	1.	0.89	0.11
12	7.	4.	3.	0.57	0.43
13	1.	0.	1.	0.	1.00
14	2.	1.	1.	0.50	0.50
15	0.	0.	0.	0.	0.
TOTAL	866.	633.	233.	0.73	0.27

	MEAN	S.D.
CORRECT DOCUMENTS	5.45	2.28
INCORRECT DOCUMENTS	4.87	2,35

TABLE 1 EFFECTIVENESS VS NUMBER OF SENTENCES
140 DOCUMENTS IN EACH CATEGORY   DISCRIMINATING WORD SET   2   48 DISCRIMINATING WORDS

TEST DOCUMENTS			CATEGORY TOTAL		
NUMBER OF SENTENCES	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
1	23.	13.	10.	0.57	0.43
2	115.	79.	36.	0.69	0.31
3	207.	145.	62.	0.70	0.30
4	263.	202.	61.	0.77	0.23
5	232.	181.	51.	0.78	0.22
6	175.	132.	43.	0.75	0.25
7	116.	90.	26.	0.78	0.22
8	88.	70.	18.	0.80	0.20
9	60.	53.	7.	0.88	0.12
10	29.	23.	6.	0.79	0.21
11	11.	10.	1.	0.91	0.09
12	9.	6.	3.	0.67	0.33
13	1.	0.	1.	0.	1.00
14	4.	3.	1.	0.75	0.25
15	0.	0.	0.	0.	0.
TOTAL	1333.	1007.	326.	0.76	0.24

	MEAN	S.D.
CORRECT DOCUMENTS	5.20	2.25
INCORRECT DOCUMENTS	4.75	2.26

TABLE 2 EFFECTIVENESS VS DOCUMENT LENGTH
140 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 2 48 DISCRIMINATING WORDS

TEST DOCUMENTS			CATEGORY A		
NUMBER OF TOKENS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0 - 9	0.	0.	0.	0.	0.
10 - 19	0.	0.	0.	0.	0.
20 - 29	3.	3.	0.	1.00	0.
30 - 39	12.	10.	2.	0.83	0.17
40 - 49	20.	19.	1.	0.95	0.05
50 - 59	25.	21.	4.	0.84	0.16
60 - 69	30.	28.	2.	0.93	0.07
70 - 79	26.	24.	2.	0.92	0.08
80 - 89	30.	27.	3.	0.90	0.10
90 - 99	24.	19.	5.	0.79	0.21
100 - 109	18.	15.	3.	0.83	0.17
110 - 119	12.	11.	1.	0.92	0.08
120 - 129	10.	9.	1.	0.90	0.10
130 - 139	2.	2.	0.	1.00	0.
140 - 149	4.	4.	0.	1.00	0.
150 - 159	2.	1.	1.	0.50	0.50
160 - 169	2.	1.	1.	0.50	0.50
170 - 179	1.	1.	0.	1.00	0.
180 - 189	1.	1.	0.	1.00	0.
190 - 199	0.	0.	0.	0.	0.
200 - 209	0.	0.	0.	0.	0.
210 - 219	0.	0.	0.	0.	0.
220 - 229	0.	0.	0.	0.	0.
230 - 239	0.	0.	0.	0.	0.
240 - 249	0.	0.	0.	0.	0.
250 - UP	0.	0.	0.	0.	0.
TOTAL	222.	196.	26.	0.88	0.12

	MEAN	S.D.
CORRECT DOCUMENTS	79.47	29.85
INCORRECT DOCUMENTS	85.31	32.04

TABLE 2 EFFECTIVENESS VS DOCUMENT LENGTH
140 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 2 48 DISCRIMINATING WORDS

TEST DOCUMENTS		CATEGORY M			
NUMBER OF TOKENS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0 - 9	0.	0.	0.	0.	0.
10 - 19	1.	0.	1.	0.	1.00
20 - 29	7.	3.	4.	0.43	0.57
30 - 39	15.	11.	4.	0.73	0.27
40 - 49	17.	12.	5.	0.71	0.29
50 - 59	23.	16.	7.	0.70	0.30
60 - 69	32.	21.	11.	0.66	0.34
70 - 79	21.	14.	7.	0.67	0.33
80 - 89	23.	17.	6.	0.74	0.26
90 - 99	16.	12.	4.	0.75	0.25
100 - 109	19.	15.	4.	0.79	0.21
110 - 119	11.	7.	4.	0.64	0.36
120 - 129	19.	15.	4.	0.79	0.21
130 - 139	7.	7.	0.	1.00	0.
140 - 149	6.	6.	0.	1.00	0.
150 - 159	5.	2.	3.	0.40	0.60
160 - 169	4.	4.	0.	1.00	0.
170 - 179	5.	5.	0.	1.00	0.
180 - 189	5.	3.	2.	0.60	0.40
190 - 199	2.	2.	0.	1.00	0.
200 - 209	4.	3.	1.	0.75	0.25
210 - 219	1.	1.	0.	1.00	0.
220 - 229	1.	1.	0.	1.00	0.
230 - 239	0.	0.	0.	0.	0.
240 - 249	1.	1.	0.	1.00	0.
250 - UP	0.	0.	0.	0.	0.
TOTAL	245.	178.	67.	0.73	0.27

	MEAN	S.D.
CORRECT DOCUMENTS	95.60	45.63
INCORRECT DOCUMENTS	80.31	39.30

TABLE 2 EFFECTIVENESS VS DOCUMENT LENGTH
140 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 2 48 DISCRIMINATING WORDS

TEST DOCUMENTS			CATEGORY P		
NUMBER OF TOKENS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0 - 9	0.	0.	0.	0.	0.
10 - 19	0.	0.	0.	0.	0.
20 - 29	8.	4.	4.	0.50	0.50
30 - 39	31.	18.	13.	0.58	0.42
40 - 49	44.	30.	14.	0.68	0.32
50 - 59	64.	37.	27.	0.58	0.42
60 - 69	82.	55.	27.	0.67	0.33
70 - 79	80.	56.	24.	0.70	0.30
80 - 89	66.	48.	18.	0.73	0.27
90 - 99	68.	54.	14.	0.79	0.21
100 - 109	60.	47.	13.	0.78	0.22
110 - 119	59.	43.	16.	0.73	0.27
120 - 129	71.	52.	19.	0.73	0.27
130 - 139	46.	35.	11.	0.76	0.24
140 - 149	45.	32.	13.	0.71	0.29
150 - 159	28.	23.	5.	0.82	0.18
160 - 169	29.	26.	3.	0.90	0.10
170 - 179	26.	21.	5.	0.81	0.19
180 - 189	18.	17.	1.	0.94	0.06
190 - 199	14.	14.	0.	1.00	0.
200 - 209	10.	8.	2.	0.80	0.20
210 - 219	4.	2.	2.	0.50	0.50
220 - 229	4.	4.	0.	1.00	0.
230 - 239	6.	5.	1.	0.83	0.17
240 - 249	1.	1.	0.	1.00	0.
250 - UP	2.	1.	1.	0.50	0.50
TOTAL	866.	633.	233.	0.73	0.27

	MEAN	S.D.
CORRECT DOCUMENTS	108.36	45.94
INCORRECT DOCUMENTS	92.71	42.75

TABLE 2   EFFECTIVENESS VS DOCUMENT LENGTH
140 DOCUMENTS IN EACH CATEGORY   DISCRIMINATING WORD SET   2   48 DISCRIMINATING WORDS

TEST DOCUMENTS			CATEGORY TOTAL		
NUMBER OF TOKENS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0 - 9	0.	0.	0.	0.	0.
10 - 19	1.	0.	1.	0.	1.00
20 - 29	18.	10.	8.	0.56	0.44
30 - 39	58.	39.	19.	0.67	0.33
40 - 49	81.	61.	20.	0.75	0.25
50 - 59	112.	74.	38.	0.66	0.34
60 - 69	144.	104.	40.	0.72	0.28
70 - 79	127.	94.	33.	0.74	0.26
80 - 89	119.	92.	27.	0.77	0.23
90 - 99	108.	85.	23.	0.79	0.21
100 - 109	97.	77.	20.	0.79	0.21
110 - 119	82.	61.	21.	0.74	0.26
120 - 129	100.	76.	24.	0.76	0.24
130 - 139	55.	44.	11.	0.80	0.20
140 - 149	55.	42.	13.	0.76	0.24
150 - 159	35.	26.	9.	0.74	0.26
160 - 169	35.	31.	4.	0.89	0.11
170 - 179	32.	27.	5.	0.84	0.16
180 - 189	24.	21.	3.	0.88	0.13
190 - 199	16.	16.	0.	1.00	0.
200 - 209	14.	11.	3.	0.79	0.21
210 - 219	5.	3.	2.	0.60	0.40
220 - 229	5.	5.	0.	1.00	0.
230 - 239	6.	5.	1.	0.83	0.17
240 - 249	2.	2.	0.	1.00	0.
250 - UP	2.	1.	1.	0.50	0.50
TOTAL	1333.	1007.	326.	0.76	0.24

	MEAN	S.D.
CORRECT DOCUMENTS	100.48	44.69
INCORRECT DOCUMENTS	89.57	41.61

TABLE 3 EFFECTIVENESS VS WORDS IN INTERSECTION OF DOCUMENTS AND DISCRIMINATING SET
140 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 2 48 DISCRIMINATING WORDS

TEST DOCUMENTS			CATEGORY A		
NUMBER OF WORDS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0	1.	0.	1.	0.	1.00
1	1.	1.	0.	1.00	0.
2	4.	2.	2.	0.50	0.50
3	16.	14.	2.	0.8F	0.13
4	29.	23.	6.	0.79	0.21
5	27.	23.	4.	0.85	0.15
6	35.	32.	3.	0.91	0.09
7	35.	32.	3.	0.91	0.09
8	28.	26.	2.	0.93	0.07
9	23.	20.	3.	0.87	0.13
10	10.	10.	0.	1.00	0.
11	8.	8.	0.	1.00	0.
12	3.	3.	0.	1.00	0.
13	1.	1.	0.	1.00	0.
14	1.	1.	0.	1.00	0.
15	0.	0.	0.	0.	0.
16	0.	0.	0.	0.	0.
17	0.	0.	0.	0.	0.
18	0.	0.	0.	0.	0.
19	0.	0.	0.	0.	0.
20	0.	0.	0.	0.	0.
TOTAL	222.	196.	26.	0.88	0.12

	MEAN	S.D.
CORRECT DOCUMENTS	6.68	2.37
INCORRECT DOCUMENTS	5.23	2.29

TABLE 3   EFFECTIVENESS VS WORDS IN INTERSECTION OF DOCUMENTS AND DISCRIMINATING SET
140 DOCUMENTS IN EACH CATEGORY   DISCRIMINATING WORD SET   2   48 DISCRIMINATING WORDS

TEST DOCUMENTS		CATEGORY M			
NUMBER OF WORDS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0	6.	0.	6.	0.	1.00
1	3.	2.	1.	0.67	0.33
2	14.	10.	4.	0.71	0.29
3	26.	14.	12.	0.54	0.46
4	20.	12.	8.	0.60	0.40
5	25.	19.	6.	0.76	0.24
6	37.	31.	6.	0.84	0.16
7	39.	28.	11.	0.72	0.28
8	30.	24.	6.	0.80	0.20
9	22.	19.	3.	0.86	0.14
10	7.	5.	2.	0.71	0.29
11	7.	6.	1.	0.86	0.14
12	5.	4.	1.	0.80	0.20
13	2.	2.	0.	1.00	0.
14	0.	0.	0.	0.	0.
15	0.	0.	0.	0.	0.
16	1.	1.	0.	1.00	0.
17	0.	0.	0.	0.	0.
18	0.	0.	0.	0.	0.
19	1.	1.	0.	1.00	0.
20	0.	0.	0.	0.	0.
TOTAL	245.	178.	67.	0.73	0.27

	MEAN	S.D.
CORRECT DOCUMENTS	6.61	2.78
INCORRECT DOCUMENTS	5.54	2.85

TABLE 3 EFFECTIVENESS VS WORDS IN INTERSECTION OF DOCUMENTS AND DISCRIMINATING SET
140 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 2 48 DISCRIMINATING WORDS

TEST DOCUMENTS			CATEGORY P		
NUMBER OF WORDS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0	2.	2.	0.	1.00	0.
1	11.	5.	6.	0.45	0.55
2	19.	8.	11.	0.42	0.58
3	34.	17.	17.	0.50	0.50
4	66.	39.	27.	0.59	0.41
5	96.	62.	36.	0.63	0.37
6	101.	67.	34.	0.66	0.34
7	126.	88.	38.	0.70	0.30
8	92.	74.	18.	0.80	0.20
9	92.	74.	18.	0.80	0.20
10	71.	59.	12.	0.83	0.17
11	44.	36.	8.	0.82	0.18
12	46.	42.	4.	0.91	0.09
13	30.	27.	3.	0.90	0.10
14	15.	15.	0.	1.00	0.
15	8.	8.	0.	1.00	0.
16	5.	5.	0.	1.00	0.
17	3.	2.	1.	0.67	0.33
18	2.	2.	0.	1.00	0.
19	1.	1.	0.	1.00	0.
20	0.	0.	0.	0.	0.
TOTAL	866.	633,	233.	0.73	0.27

	MEAN	S.D.
CORRECT DOCUMENTS	8.14	3.16
INCORRECT DOCUMENTS	6.24	2.69

TABLE 3 EFFECTIVENESS VS WORDS IN INTERSECTION OF DOCUMENTS AND DISCRIMINATING SET
140 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 2 48 DISCRIMINATING WORDS

TEST DOCUMENTS			CATEGORY TOTAL		
NUMBER OF WORDS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0	9.	2.	7.	0.22	0.78
1	15.	8.	7.	0.53	0.47
2	37.	20.	17.	0.54	0.46
3	76.	45.	31.	0.59	0.41
4	115.	74.	41.	0.64	0.36
5	150.	104.	46.	0.69	0.31
6	173.	130.	43.	0.75	0.25
7	200.	148.	52.	0.74	0.26
8	150.	124.	26.	0.83	0.17
9	137.	113.	24.	0.82	0.18
10	88.	74.	14.	0.84	0.16
11	59.	50.	9.	0.85	0.15
12	54.	49.	5.	0.91	0.09
13	33.	39.	3.	0.91	0.09
14	16.	16.	0.	1.00	0.
15	8.	8.	0.	1.00	0.
16	6.	6.	0.	1.00	0.
17	3.	2.	1.	0.67	0.33
18	2.	2.	0.	1.00	0.
19	2.	2.	0.	1.00	0.
20	0.	0.	0.	0.	0.
TOTAL	1333.	1007.	326.	0.76	0.24

	MEAN	S.D.
CORRECT DOCUMENTS	7.59	3.04
INCORRECT DOCUMENTS	5.92	2.74

TABLE 4 EFFECTIVENESS VS RADIUS
140 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 2 48 DISCRIMINATING WORDS

TEST DOCUMENTS			CATEGORY A		
RADIUS	NUMBER OF DUCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0. - 0.49	49.	34.	15.	0.69	0.31
0.5 - 0.99	66.	60.	6.	0.91	0.09
1.0 - 1.99	83.	78.	5.	0.94	0.06
2.0 - 2.99	18.	18.	0.	1.00	0.
3.0 - 3.99	6.	6.	0.	1.00	0.
4.0 - 4.99	0.	0.	0.	0.	0.
5.0 - 5.99	0.	0.	0.	0.	0.
6.0 - 6.99	0.	0.	0.	0.	0.
7.0 - 7.99	0.	0.	0.	0.	0.
8.0 - 8.99	0.	0.	0.	0.	0.
9.0 - 9.99	0.	0.	0.	0.	0.
10. - UP	0.	0.	0.	0.	0.
TOTAL	222.	196.	26.	0.88	0.12

	MEAN	S.D.
CORRECT DOCUMENTS	1.17	0.71
INCORRECT DOCUMENTS	0.51	0.39

TABLE 4 EFFECTIVENESS VS RADIUS
140 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 2 48 DISCRIMINATING WORDS

TEST DOCUMENTS		CATEGORY M			
RADIUS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0. - 0.49	79.	40.	39.	0.51	0.49
0.5 - 0.99	80.	60.	20.	0.75	0.25
1.0 - 1.99	70.	63.	7.	0.90	0.10
2.0 - 2.99	13.	12.	1.	0.92	0.08
3.0 - 3.99	3.	3.	0.	1.00	0.
4.0 - 4.99	0.	0.	0.	0.	0.
5.0 - 5.99	0.	0.	0.	0.	0.
6.0 - 6.99	0.	0.	0.	0.	0.
7.0 - 7.99	0.	0.	0.	0.	0.
8.0 - 8.99	0.	0.	0.	0.	0.
9.0 - 9.99	0.	0.	0.	0.	0.
10. - UP	0.	0.	0.	0.	0.
TOTAL	245.	178.	67.	0.73	0.27

	MEAN	S.D.
CORRECT DOCUMENTS	1.04	0.68
INCORRECT DOCUMENTS	0.54	0.50

TABLE 4 EFFECTIVENESS VS RADIUS
140 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 2 43 DISCRIMINATING WORDS

TEST DOCUMENTS			CATEGORY P		
RADIUS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0. - 0.49	203.	106.	97.	0.52	0.48
0.5 - 0.99	310.	211.	99.	0.68	0.32
1.0 - 1.99	272.	238.	34.	0.88	0.13
2.0 - 2.99	69.	66.	3.	0.96	0.04
3.0 - 3.99	9.	9.	0.	1.00	0.
4.0 - 4.99	3.	3.	0.	1.00	0.
5.0 - 5.99	0.	0.	0.	0.	0.
6.0 - 6.99	0.	0.	0.	0.	0.
7.0 - 7.99	0.	0.	0.	0.	0.
8.0 - 8.99	0.	0.	0.	0.	0.
9.0 - 9.99	0.	0.	0.	0.	0.
10. - UP	0.	0.	0.	0.	0.
TOTAL	866.	633.	233.	0.73	0.27

	MEAN	S.D.
CORRECT DOCUMENTS	1.13	0.72
INCORRECT DOCUMENTS	0.66	0.43

TABLE 4 EFFECTIVENESS VS RADIUS
140 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 2 48 DISCRIMINATING WORDS

	TEST DOCUMENTS		CATEGORY TOTAL		
RADIUS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
C. - 0.49	331.	180.	151.	0.54	0.46
0.5 - 0.99	456.	331.	125.	0.73	0.27
1.0 - 1.99	425.	379.	46.	0.89	0.11
2.0 - 2.99	100.	96.	4.	0.96	0.04
3.0 - 3.99	18.	18.	0.	1.00	C.
4.0 - 4.99	3.	3.	0.	1.00	0.
5.0 - 5.99	0.	0.	0.	0.	0.
6.0 - 6.99	0.	0.	0.	0.	0.
7.0 - 7.99	0.	0.	0.	0.	0.
8.0 - 8.99	0.	0.	0.	0.	0.
9.0 - 9.99	0.	0.	0.	0.	0.
10. - UP	0.	0.	0.	0.	0.
TOTAL	333.	1007.	326.	0.76	0.24

	MEAN	S.D.
CORRECT DOCUMENTS	1.12	0.71
INCORRECT DOCUMENTS	0.62	0.45

TABLE 5 DOCUMENT CLASSIFICATION SUMMARY
140 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 2 48 DISCRIMINATING WORDS

TEST

ACTUAL CATEGORY	AUTO CATEGORY A	M	P	TOTAL
A	196.00	10.00	16.00	222.00
M	15.00	178.00	52.00	245.00
P	73.00	160.00	633.00	866.00
TOTAL	284.00	348.00	701.00	1333.00
	PERCENTAGE			
A	0.88	0.05	0.07	1.00
M	0.06	0.73	0.21	1.00
P	0.08	0.18	0.73	1.00

	SWETS MEASURES				
	PERTINENT	NOT PERTINENT	RECALL RATIO	RELEVANCE RATIO	PRECISION RATIO
CATEGORY A RETRIEVED	0.88	0.06	0.88	0.69	1.28
CATEGORY M RETRIEVED	0.73	0.11	0.73	0.51	1.42
CATEGORY P RETRIEVED	0.73	0.08	0.73	0.90	0.81

TABLE 2 EFFECTIVENESS VS NUMBER OF SENTENCES
35 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 7 48 DISCRIMINATING WORDS

SAMPLE DOCUMENTS			CATEGORY P3		
NUMBER OF SENTENCES	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
1	1.	1.	0.	1.00	0.
2	4.	4.	0.	1.00	0.
3	5.	5.	0.	1.00	0.
4	9.	9.	0.	1.00	0.
5	2.	2.	0.	1.00	0.
6	3.	3.	0.	1.00	0.
7	2.	2.	0.	1.00	0.
8	3.	3.	0.	1.00	0.
9	3.	3.	0.	1.00	0.
10	2.	2.	0.	1.00	0.
11	1.	1.	0.	1.00	0.
12	0.	0.	0.	0.	0.
13	0.	0.	0.	0.	0.
14	0.	0.	0.	0.	0.
15	0.	0.	0.	0.	0.
TOTAL	35.	35.	0.	1.00	0.

	MEAN	S.D.
CORRECT DOCUMENTS	5.26	2.68
INCORRECT DOCUMENTS	0.	0.

TABLE 1 EFFECTIVENESS VS NUMBER OF SENTENCES
35 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 7 48 DISCRIMINATING WORDS

SAMPLE DOCUMENTS			CATEGORY P4		
NUMBER OF SENTENCES	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
1	0.	0.	0.	0.	0.
2	3.	2.	1.	0.67	0.33
3	7.	6.	1.	0.86	0.14
4	6.	6.	0.	1.00	0.
5	7.	7.	0.	1.00	0.
6	1.	1.	0.	1.00	0.
7	6.	6.	0.	1.00	0.
8	2.	2.	0.	1.00	0.
9	1.	1.	0.	1.00	0.
10	2.	2.	0.	1.00	0.
11	0.	0.	0.	0.	0.
12	0.	0.	0.	0.	0.
13	0.	0.	0.	0.	0.
14	0.	0.	0.	0.	0.
15	0.	0.	0.	0.	0.
TOTAL	35.	33.	2.	0.94	0.06

	MEAN	S.D.
CORRECT DOCUMENTS	5.27	2.16
INCORRECT DOCUMENTS	2.50	0.50

TABLE 1 EFFECTIVENESS VS NUMBER OF SENTENCES
35 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 7 48 DISCRIMINATING WORDS

SAMPLE DOCUMENTS			CATEGORY TOTAL		
NUMBER OF SENTENCES	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
1	1.	1.	0.	1.00	0.
2	11.	9.	2.	0.82	0.18
3	18.	16.	2.	0.89	0.11
4	18.	18.	0.	1.00	0.
5	17.	17.	0.	1.00	0.
6	10.	10.	0.	1.00	0.
7	12.	12.	0.	1.00	0.
8	5.	5.	0.	1.00	0.
9	7.	7.	0.	1.00	0.
10	4.	4.	0.	1.00	0.
11	2.	2.	0.	1.00	0.
12	0.	0.	0.	0.	0.
13	0.	0.	0.	0.	0.
14	0.	0.	0.	0.	0.
15	0.	0.	0.	0.	0.
TOTAL	105.	101.	4.	0.96	0.04

	MEAN	S.D.
CORRECT DOCUMENTS	5.28	2.35
INCORRECT DOCUMENTS	2.50	0.50

```
TABLE 2    EFFECTIVENESS VS DOCUMENT LENGTH
           35 DOCUMENTS IN EACH CATEGORY    DISCRIMINATING WORD SET     7      48 DISCRIMINATING  WORDS

   SAMPLE DOCUMENTS                                  CATEGORY P2
NUMBER OF      NUMBER OF     NUMBER OF                  NUMBER OF                   PERCENTAGE OF        PERCENTAGE OF
 TOKENS        DOCUMENTS      CORRECT DOCUMENTS      INCORRECT DOCUMENTS          CORRECT DOCUMENTS    INCORRECT DOCUMENTS
   0 -    9     0.                0.                      0.                         0.                   0.
  10 -   19     0.                0.                      0.                         0.                   0.
  20 -   29     0.                0.                      0.                         0.                   0.
  30 -   39     1.                1.                      0.                         1.00                 0.
  40 -   49     3.                1.                      2.                         0.33                 0.67
  50 -   59     4.                4.                      0.                         1.00                 0.
  60 -   69     3.                3.                      0.                         1.00                 0.
  70 -   79     3.                3.                      0.                         1.00                 0.
  80 -   89     5.                5.                      0.                         1.00                 0.
  90 -   99     1.                1.                      0.                         1.00                 0.
 100 -  109     1.                1.                      0.                         1.00                 0.
 110 -  119     1.                1.                      0.                         1.00                 0.
 120 -  129     3.                3.                      0.                         1.00                 0.
 130 -  139     0.                0.                      0.                         0.                   0.
 140 -  149     2.                2.                      0.                         1.00                 0.
 150 -  159     1.                1.                      0.                         1.00                 0.
 160 -  169     1.                1.                      0.                         1.00                 0.
 170 -  179     2.                2.                      0.                         1.00                 0.
 180 -  189     1.                1.                      0.                         1.00                 0.
 190 -  199     2.                2.                      0.                         1.00                 0.
 200 -  209     1.                1.                      0.                         1.00                 0.
 210 -  219     0.                0.                      0.                         0.                   0.
 220 -  229     0.                0.                      0.                         0.                   0.
 230 -  239     0.                0.                      0.                         0.                   0.
 240 -  249     0.                0.                      0.                         0.                   0.
 250 -   UP     0.                0.                      0.                         0.                   0.

 TOTAL         35.               33.                      2.                         0.94                 0.06

                            MEAN              S.D.
 CORRECT   DOCUMENTS       107.94             49.78
 INCORRECT DOCUMENTS        45.50              0.50
```

TABLE 2   EFFECTIVENESS VS DOCUMENT LENGTH
35 DOCUMENTS IN EACH CATEGORY   DISCRIMINATING WORD SET   7   48 DISCRIMINATING WORDS

SAMPLE DOCUMENTS NUMBER OF TOKENS	 NUMBER OF DOCUMENTS	 NUMBER OF CORRECT DOCUMENTS	CATEGORY P3 NUMBER OF INCORRECT DOCUMENTS	 PERCENTAGE OF CORRECT DOCUMENTS	 PERCENTAGE OF INCORRECT DOCUMENTS
0 - 9	0.	0.	0.	0.	0.
10 - 19	0.	0.	0.	0.	0.
20 - 29	1.	1.	0.	1.00	0.
30 - 39	3.	3.	0.	1.00	0.
40 - 49	2.	2.	0.	1.00	0.
50 - 59	2.	2.	0.	1.00	0.
60 - 69	4.	4.	0.	1.00	0.
70 - 79	1.	1.	0.	1.00	0.
80 - 89	2.	2.	0.	1.00	0.
90 - 99	3.	3.	0.	1.00	0.
100 - 109	3.	3.	0.	1.00	0.
110 - 119	2.	2.	0.	1.00	0.
120 - 129	1.	1.	0.	1.00	0.
130 - 139	2.	2.	0.	1.00	0.
140 - 149	2.	2.	0.	1.00	0.
150 - 159	2.	2.	0.	1.00	0.
160 - 169	2.	2.	0.	1.00	0.
170 - 179	2.	2.	0.	1.00	0.
180 - 189	1.	1.	0.	1.00	0.
190 - 199	0.	0.	0.	0.	0.
200 - 209	0.	0.	0.	0.	0.
210 - 219	0.	0.	0.	0.	0.
220 - 229	0.	0.	0.	0.	0.
230 - 239	0.	0.	0.	0.	0.
240 - 249	0.	0.	0.	0.	0.
250 - UP	0.	0.	0.	0.	0.
TOTAL	35.	35.	0.	1.00	0.

	MEAN	S.D.
CORRECT DOCUMENTS	100.14	46.05
INCORRECT DOCUMENTS	0.	0.

TABLE 2 EFFECTIVENESS VS DOCUMENT LENGTH
35 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 7 48 DISCRIMINATING WORDS

SAMPLE DOCUMENTS			CATEGORY P4		
NUMBER OF TOKENS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0 - 9	0.	0.	0.	0.	0.
10 - 19	0.	0.	0.	0.	0.
20 - 29	0.	0.	0.	0.	0.
30 - 39	2.	1.	1.	0.50	0.50
40 - 49	2.	2.	0.	1.00	0.
50 - 59	2.	2.	0.	1.00	0.
60 - 69	5.	4.	1.	0.80	0.20
70 - 79	4.	4.	0.	1.00	0.
80 - 89	3.	3.	0.	1.00	0.
90 - 99	2.	2.	0.	1.00	0.
100 - 109	5.	5.	0.	1.00	0.
110 - 119	2.	2.	0.	1.00	0.
120 - 129	.	4.	0.	1.00	0.
130 - 139	1.	1.	0.	1.00	0.
140 - 149	1.	1.	0.	1.00	0.
150 - 159	1.	1.	0.	1.00	0.
160 - 169	1.	1.	0.	1.00	0.
170 - 179	0.	0.	0.	0.	0.
180 - 189	0.	0.	0.	0.	0.
190 - 199	0.	0.	0.	0.	0.
200 - 209	0.	0.	0.	0.	0.
210 - 219	0.	0.	0.	0.	0.
220 - 229	0.	0.	0.	0.	0.
230 - 239	0.	0.	0.	0.	0.
240 - 249	0.	0.	0.	0.	0.
250 - UP	0.	0.	0.	0.	0.
TOTAL	35.	33.	2.	0.94	0.06

	MEAN	S.D.
CORRECT DOCUMENTS	93.85	33.15
INCORRECT DOCUMENTS	45.00	15.00

TABLE 2 EFFECTIVENESS VS DOCUMENT LENGTH

35 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 7 48 DISCRIMINATING WORDS

SAMPLE DOCUMENTS			CATEGORY TOTAL		
NUMBER OF TOKENS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0 - 9	0.	0.	0.	0.	0.
10 - 19	0.	0.	0.	0.	0.
20 - 29	1.	1.	0.	1.00	0.
30 - 39	6.	5.	1.	0.83	0.17
40 - 49	7.	5.	2.	0.71	0.29
50 - 59	8.	8.	0.	1.00	0.
60 - 69	12.	11.	1.	0.92	0.08
70 - 79	8.	8.	0.	1.00	0.
80 - 89	10.	10.	0.	1.00	0.
90 - 99	6.	6.	0.	1.00	0.
100 - 109	9.	9.	0.	1.00	0.
110 - 119	5.	5.	0.	1.00	0.
120 - 129	8.	8.	0.	1.00	0.
130 - 139	3.	3.	0.	1.00	0.
140 - 149	5.	5.	0.	1.00	0.
150 - 159	4.	4.	0.	1.00	0.
160 - 169	4.	4.	0.	1.00	0.
170 - 179	4.	4.	0.	1.00	0.
180 - 189	2.	2.	0.	1.00	0.
190 - 199	2.	2.	0.	1.00	0.
200 - 209	1.	1.	0.	1.00	0.
210 - 219	0.	0.	0.	0.	0.
220 - 229	0.	0.	0.	0.	0.
230 - 239	0.	0.	0.	0.	0.
240 - 249	0.	0.	0.	0.	0.
250 - UP	0.	0.	0.	0.	0.
TOTAL	105.	101.	4.	0.96	0.04

	MEAN	S.D.
CORRECT DOCUMENTS	100.63	44.00
INCORRECT DOCUMENTS	45.25	10.62

TABLE 3 EFFECTIVENESS VS WORDS IN INTERSECTION OF DOCUMENTS AND DISCRIMINATING SET
35 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 7 48 DISCRIMINATING WORDS

SAMPLE DOCUMENTS			CATEGORY P2		
NUMBER OF WORDS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0	1.	0.	1.	0.	1.00
1	3.	2.	1.	0.67	0.33
2	4.	4.	0.	1.00	0.
3	4.	4.	0.	1.00	0.
4	6.	6.	0.	1.00	0.
5	7.	7.	0.	1.00	0.
6	5.	5.	0.	1.00	0.
7	3.	3.	0.	1.00	0.
8	2.	2.	0.	1.00	0.
9	0.	0.	0.	0.	0.
10	0.	0.	0.	0.	0.
11	0.	0.	0.	0.	0.
12	0.	0.	0.	0.	0.
13	0.	0.	0.	0.	0.
14	0.	0.	0.	0.	0.
15	0.	0.	0.	0.	0.
16	0.	0.	0.	0.	0.
17	0.	0.	0.	0.	0.
18	0.	0.	0.	0.	0.
19	0.	0.	0.	0.	0.
20	0.	0.	0.	0.	0.
TOTAL	35.	33.	2.	0.94	0.06

	MEAN	S.D.
CORRECT DOCUMENTS	4.48	1.88
INCORRECT DOCUMENTS	0.50	0.50

TABLE 3 EFFECTIVENESS VS WORDS IN INTERSECTION OF DOCUMENTS AND DISCRIMINATING SET
35 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 7 48 DISCRIMINATING WORDS

SAMPLE DOCUMENTS			CATEGORY P3		
NUMBER OF WORDS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0	0.	0.	0.	0.	0.
1	1.	1.	0.	1.00	0.
2	3.	3.	0.	1.00	0.
3	5.	5.	0.	1.00	0.
4	4.	4.	0.	1.00	0.
5	6.	6.	0.	1.00	0.
6	4.	4.	0.	1.00	0.
7	4.	4.	0.	1.00	0.
8	4.	4.	0.	1.00	0.
9	3.	3.	0.	1.00	0.
10	1.	1.	0.	1.00	0.
11	0.	0.	0.	0.	0.
12	0.	0.	0.	0.	0.
13	0.	0.	0.	0.	0.
14	0.	0.	0.	0.	0.
15	0.	0.	0.	0.	0.
16	0.	0.	0.	0.	0.
17	0.	0.	0.	0.	0.
18	0.	0.	0.	0.	0.
19	0.	0.	0.	0.	0.
20	0.	0.	0.	0.	0.
TOTAL	35.	35.	0.	1.00	0.

	MEAN	S.D.
CORRECT DOCUMENTS	5.40	2.33
INCORRECT DOCUMENTS	0.	0.

TABLE 3   EFFECTIVENESS VS WORDS IN INTERSECTION OF DOCUMENTS AND DISCRIMINATING SET
35 DOCUMENTS IN EACH CATEGORY   DISCRIMINATING WORD SET   7   48 DISCRIMINATING WORDS

SAMPLE DOCUMENTS		CATEGORY P4			
NUMBER OF WORDS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0	0.	0.	0.	0.	0.
1	3.	2.	1.	0.67	0.33
2	3.	2.	1.	0.67	0.33
3	2.	2.	0.	1.00	0.
4	6.	6.	0.	1.00	0.
5	4.	4.	0.	1.00	0.
6	6.	6.	0.	1.00	0.
7	5.	5.	0.	1.00	0.
8	1.	1.	0.	1.00	0.
9	2.	2.	0.	1.00	0.
10	0.	0.	0.	0.	0.
11	3.	3.	0.	1.00	0.
12	0.	0.	0.	0.	0.
13	0.	0.	0.	0.	0.
14	0.	0.	0.	0.	0.
15	0.	0.	0.	0.	0.
16	0.	0.	0.	0.	0.
17	0.	0.	0.	0.	0.
18	0.	0.	0.	0.	0.
19	0.	0.	0.	0.	0.
20	0.	0.	0.	0.	0.
TOTAL	35.	33.	2.	0.94	0.06

	MEAN	S.D.
CORRECT DOCUMENTS	5.64	2.61
INCORRECT DOCUMENTS	1.50	0.50

TABLE 3 EFFECTIVENESS VS WORDS IN INTERSECTION OF DOCUMENTS AND DISCRIMINATING SET
35 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 7 48 DISCRIMINATING WORDS

SAMPLE DOCUMENTS			CATEGORY TOTAL		
NUMBER OF WORDS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0	1.	0.	1.	0.	1.00
1	7.	5.	2.	0.71	0.29
2	10.	9.	1.	0.90	0.10
3	11.	11.	0.	1.00	0.
4	16.	16.	0.	1.00	0.
5	17.	17.	0.	1.00	0.
6	15.	15.	0.	1.00	0.
7	12.	12.	0.	1.00	0.
8	7.	7.	0.	1.00	0.
9	5.	5.	0.	1.00	0.
10	1.	1.	0.	1.00	0.
11	3.	3.	0.	1.00	0.
12	0.	0.	0.	0.	0.
13	0.	0.	0.	0.	0.
14	0.	0.	0.	0.	0.
15	0.	0.	0.	0.	0.
16	0.	0.	0.	0.	0.
17	0.	0.	0.	0.	0.
18	0.	0.	0.	0.	0.
19	0.	0.	0.	0.	0.
20	0.	0.	0.	0.	0.
TOTAL	105.	101.	4.	0.96	0.04

	MEAN	S.D.
CORRECT DOCUMENTS	5.18	2.35
INCORRECT DOCUMENTS	1.00	0.71

TABLE 4 EFFECTIVENESS VS RADIUS
35 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 7 48 DISCRIMINATING WORDS

SAMPLE DOCUMENTS			CATEGORY P2		
RADIUS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0. - 0.49	20.	18.	2.	0.90	0.10
0.5 - 0.99	15.	15.	0.	1.00	0.
1.0 - 1.99	0.	0.	0.	0.	0.
2.0 2.99	0.	0.	0.	0.	0.
3.0 - 3.99	0.	0.	0.	0.	0.
4.0 - 4.99	0.	0.	0.	0.	0.
5.0 - 5.99	0.	0.	0.	0.	0.
6.0 - 6.99	0.	0.	0.	0.	0.
7.0 - 7.99	0.	0.	0.	0.	0.
8.0 - 8.99	0.	0.	0.	0.	0.
9.0 - 9.99	0.	0.	0.	0.	0.
10. - UP	0.	0.	0.	0.	0.
TOTAL	35.	33.	2.	0.94	0.06

	MEAN	S.D.
CORRECT DOCUMENTS	0.44	0.15
INCORRECT DOCUMENTS	0.04	0.04

```
TABLE 4   EFFECTIVENESS VS RADIUS
          35 DOCUMENTS IN EACH CATEGORY   DISCRIMINATING WORD SET    7     48 DISCRIMINATING  WORDS

   SAMPLE DOCUMENTS                               CATEGORY P3
RADIUS        NUMBER OF    NUMBER OF                NUMBER OF                 PERCENTAGE OF        PERCENTAGE OF
              DOCUMENTS     CORRECT DOCUMENTS     INCORRECT DOCUMENTS       CORRECT DOCUMENTS    INCORRECT DOCUMENTS
0.   - 0.49   10.              10.                       0.                     1.00                  0.

0.5 - 0.99    25.              25.                       0.                     1.00                  0.

1.0 - 1.99     0.               0.                       0.                     0.                    0.

2.0 - 2.99     0.               0.                       0.                     0.                    0.

3.0 - 3.99     0.               0.                       0.                     0.                    0.

4.0 - 4.99     0.               0.                       0.                     0.                    0.

5.0 - 5.99     0.               0.                       0.                     0.                    0.

6.0 - 6.99     0.               0.                       0.                     0.                    0.

7.0 - 7.99     0.               0.                       0.                     0.                    0.

8.0 - 8.99     0.               0.                       0.                     0.                    0.

9.0 - 9.99     0.               0.                       0.                     0.                    0.

10. -     UP   0.               0.                       0.                     0.                    0.

TOTAL         35.              35.                       0.                     1.00                  0.

                            MEAN              S.D.
CORRECT   DOCUMENTS          0.60              0.16
INCORRECT DOCUMENTS          0.                0.
```

TABLE 4 EFFECTIVENESS VS RADIUS
35 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 7 48 DISCRIMINATING WORDS

SAMPLE DOCUMENTS RADIUS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	CATEGORY P4 NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0. - 0.49	22.	20.	2.	0.91	0.09
0.5 - 0.99	13.	13.	0.	1.00	0.
1.0 - 1.99	0.	0.	0.	0.	0.
2.0 - 2.99	0.	0.	0.	0.	0.
3.0 - 3.99	0.	0.	0.	0.	0.
4.0 - 4.99	0.	0.	0.	0.	0.
5.0 - 5.99	0.	0.	0.	0.	0.
6.0 - 6.99	0.	0.	0.	0.	0.
7.0 - 7.99	0.	0.	0.	0.	0.
8.0 - 8.99	0.	0.	0.	0.	0.
9.0 - 9.99	0.	0.	0.	0.	0.
10. - UP	0.	0.	0.	0.	0.
TOTAL	35.	33.	2.	0.94	0.06

	MEAN	S.D.
CORRECT DOCUMENTS	0.38	0.21
INCORRECT DOCUMENTS	0.21	0.02

TABLE 4 EFFECTIVENESS VS RADIUS
35 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 7 48 DISCRIMINATING WORDS

SAMPLE DOCUMENTS			CATEGORY TOTAL		
RADIUS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0. - 0.49	52.	48.	4.	0.92	0.08
0.3 - 0.99	53.	53.	0.	1.00	0.
1.0 - 1.99	0.	0.	0.	0.	0.
2.0 - 2.99	0.	0.	0.	0.	0.
3.0 - 3.99	0.	0.	0.	0.	0.
4.0 - 4.99	0.	0.	0.	0.	0.
5.0 - 5.99	0.	0.	0.	0.	0.
6.0 - 6.99	0.	0.	0.	0.	0.
7.0 - 7.99	0.	0.	0.	0.	0.
8.0 - 8.99	0.	0.	0.	0.	0.
9.0 - 9.99	0.	0.	0.	0.	0.
10. - UP	0.	0.	0.	0.	0.
TOTAL	105.	101.	4.	0.96	0.04

	MEAN	S.D.
CORRECT DOCUMENTS	0.48	0.20
INCORRECT DOCUMENTS	0.12	0.09

```
TABLE 5   DOCUMENT CLASSIFICATION SUMMARY
          35 DOCUMENTS IN EACH CATEGORY   DISCRIMINATING WORD SET    7     48 DISCRIMINATING  WORDS

SAMPLE
                 AUTO   CATEGORY
ACTUAL
CATEGORY        P2        P3        P4       TOTAL

P2              33.00     0.        2.00     35.00

P3              0.        35.00     0.       35.00

P4              2.00      0.        33.00    35.00

TOTAL           35.00     35.00     35.00    105.00

                       PERCENTAGE

P2              0.94      0.        0.06     1.00

P3              0.        1.00      0.       1.00

P4              0.06      0.        0.94     1.00

                       SWETS MEASURES              RECALL        RELEVANCE       PRECISION
CATEGORY P2          PERTINENT   NOT PERTINENT     RATIO           RATIO           RATIO
 RETRIEVED             0.94          0.03           0.94           0.94            1.00


                       SWETS MEASURES              RECALL        RELEVANCE       PRECISION
CATEGORY P3          PERTINENT   NOT PERTINENT     RATIO           RATIO           RATIO
 RETRIEVED             1.00          0.             1.00           1.00            1.00


                       SWETS MEASURES              RECALL        RELEVANCE       PRECISION
CATEGORY P4          PERTINENT   NOT PERTINENT     RATIO           RATIO           RATIO
 RETRIEVED             0.94          0.03           0.94           0.94            1.00
```

TABLE 1 EFFECTIVENESS VS NUMBER OF SENTENCES

35 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 7 48 DISCRIMINATING WORDS

TEST DOCUMENTS		CATEGORY P2			
NUMBER OF SENTENCES	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
1	2.	0.	2.	0.	1.00
2	17.	11.	6.	0.65	0.35
3	47.	30.	17.	0.64	0.36
4	57.	42.	15.	0.74	0.26
5	51.	40.	11.	0.78	0.22
6	44.	25.	19.	0.57	0.43
7	24.	19.	5.	0.79	0.21
8	24.	19.	5.	0.79	0.21
9	11.	9.	2.	0.82	0.18
10	8.	7.	1.	0.88	0.13
11	3.	3.	0.	1.00	0.
12	2.	2.	0.	1.00	0.
13	0.	0.	0.	0.	0.
14	0.	0.	0.	0.	0.
15	0.	0.	0.	0.	0.
TOTAL	290.	207.	83.	0.71	0.29

	MEAN	S.D.
CORRECT DOCUMENTS	5.43	2.22
INCORRECT DOCUMENTS	4.78	1.92

TABLE 1 EFFECTIVENESS VS NUMBER OF SENTENCES
35 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 7 48 DISCRIMINATING WORDS

TEST DOCUMENTS		CATEGORY P3			
NUMBER OF SENTENCES	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
1	3.	0.	3.	0.	1.00
2	14.	10.	4.	0.71	0.29
3	26.	23.	3.	0.88	0.12
4	46.	38.	8.	0.83	0.17
5	41.	32.	9.	0.78	0.22
6	36.	26.	10.	0.72	0.28
7	28.	21.	7.	0.75	0.25
8	11.	10.	1.	0.91	0.09
9	15.	14.	1.	0.93	0.07
10	6.	6.	0.	1.00	0.
11	2.	2.	0.	1.00	0.
12	2.	2.	0.	1.00	0.
13	0.	0.	0.	0.	0.
14	1.	1.	0.	1.00	0.
15	0.	0.	0.	0.	0.
TOTAL	231.	185.	46.	0.80	0.20

	MEAN	S.D.
CORRECT DOCUMENTS	5.57	2.31
INCORRECT DOCUMENTS	4.85	1.90

TABLE 1 EFFECTIVENESS VS NUMBER OF SENTENCES
35 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 7 48 DISCRIMINATING WORDS

TEST DOCUMENTS			CATEGORY P4		
NUMBER OF SENTENCES	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
1	2.	2.	0.	1.00	0.
2	13.	11.	2.	0.85	0.15
3	23.	16.	7.	0.70	0.30
4	17.	12.	5.	0.71	0.29
5	15.	8.	7.	0.53	0.47
6	16.	8.	8.	0.50	0.50
7	10.	7.	3.	0.70	0.30
8	9.	6.	3.	0.67	0.33
9	1.	0.	1.	0.	1.00
10	2.	1.	1.	0.50	0.50
11	2.	2.	0.	1.00	0.
12	0.	0.	0.	0.	0.
13	0.	0.	0.	0.	0.
14	1.	1.	0.	1.00	0.
15	0.	0.	0.	0.	0.
TOTAL	111.	74.	37.	0.67	0.33

	MEAN	S.D.
CORRECT DOCUMENTS	4.74	2.52
INCORRECT DOCUMENTS	5.19	1.93

TABLE 1 EFFECTIVENESS VS NUMBER OF SENTENCES
35 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 7 48 DISCRIMINATING WORDS

TEST DOCUMENTS			CATEGORY TOTAL		
NUMBER OF SENTENCES	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
1	7.	2.	5.	0.29	0.71
2	44.	32.	12.	0.73	0.27
3	96.	69.	27.	0.72	0.28
4	120.	92.	28.	0.77	0.23
5	107.	80.	27.	0.75	0.25
6	96.	59.	37.	0.61	0.39
7	62.	47.	15.	0.76	0.24
8	44.	35.	9.	0.80	0.20
9	27.	23.	4.	0.85	0.15
10	16.	14.	2.	0.88	0.13
11	7.	7.	0.	1.00	0.
12	4.	4.	0.	1.00	0.
13	0.	0.	0.	0.	0.
14	2.	2.	0.	1.00	0.
15	0.	0.	0.	0.	0.
TOTAL	632.	466.	166.	0.74	0.26

	MEAN	S.D.
CORRECT DOCUMENTS	5.37	2.32
INCORRECT DOCUMENTS	4.89	1.92

TABLE 2 EFFECTIVENESS VS DOCUMENT LENGTH
35 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 7 48 DISCRIMINATING WORDS

TEST DOCUMENTS		CATEGORY P2			
NUMBER OF TOKENS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0 - 9	0.	0.	0.	0.	0.
10 - 19	0.	0.	0.	0.	0.
20 - 29	2.	1.	1.	0.50	0.50
30 - 39	12.	5.	7.	0.42	0.58
40 - 49	14.	10.	4.	0.71	0.29
50 - 59	22.	13.	9.	0.59	0.41
60 - 69	30.	23.	7.	0.77	0.23
70 - 79	25.	18.	7.	0.72	0.28
80 - 89	24.	18.	6.	0.75	0.25
90 - 99	22.	11.	11.	0.50	0.50
100 - 109	19.	13.	6.	0.68	0.32
110 - 119	17.	16.	1.	0.94	0.06
120 - 129	26.	22.	4.	0.85	0.15
130 - 139	18.	12.	6.	0.67	0.33
140 - 149	11.	8.	3.	0.73	0.27
150 - 159	13.	11.	2.	0.85	0.15
160 - 169	10.	7.	3.	0.70	0.30
170 - 179	11.	7.	4.	0.64	0.36
180 - 189	6.	5.	1.	0.83	0.17
190 - 199	2.	1.	1.	0.50	0.50
200 - 209	1.	1.	0.	1.00	0.
210 - 219	0.	0.	0.	0.	0.
220 - 229	1.	1.	0.	1.00	0.
230 - 239	3.	3.	0.	1.00	0.
240 - 249	0.	0.	0.	0.	0.
250 - UP	1.	1.	0.	1.00	0.
TOTAL	290.	207.	83.	0.71	0.29

	MEAN	S.D.
CORRECT DOCUMENTS	106.23	45.27
INCORRECT DOCUMENTS	94.07	42.56

TABLE 2 EFFECTIVENESS VS DOCUMENT LENGTH
35 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 7 48 DISCRIMINATING WORDS

TEST DOCUMENTS			CATEGORY P3		
NUMBER OF TOKENS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0 - 9	0.	0.	0.	0.	0.
10 - 19	0.	0.	0.	0.	0.
20 - 29	1.	0.	1.	0.	1.00
30 - 39	6.	5.	1.	0.83	0.17
40 - 49	12.	10.	2.	0.83	0.17
50 - 59	14.	9.	5.	0.64	0.36
60 - 69	18.	13.	5.	0.72	0.28
70 - 79	22.	19.	3.	0.86	0.14
80 - 89	16.	10.	6.	0.63	0.38
90 - 99	20.	17.	3.	0.85	0.15
100 - 109	18.	16.	2.	0.89	0.11
110 - 119	19.	13.	6.	0.68	0.32
120 - 129	21.	19.	2.	0.90	0.10
130 - 139	11.	7.	4.	0.64	0.36
140 - 149	7.	6.	1.	0.86	0.14
150 - 159	5.	5.	0.	1.00	0.
160 - 169	8.	6.	2.	0.75	0.25
170 - 179	10.	9.	1.	0.90	0.10
180 - 189	8.	8.	0.	1.00	0.
190 - 199	6.	5.	1.	0.83	0.17
200 - 209	3.	3.	0.	1.00	0.
210 - 219	2.	1.	1.	0.50	0.50
220 - 229	0.	0.	0.	0.	0.
230 - 239	3.	3.	0.	1.00	0.
240 - 249	1.	1.	0.	1.00	0.
250 - UP	0.	0.	0.	0.	0.
TOTAL	231.	185.	46.	0.80	0.20

	MEAN	S.D.
CORRECT DOCUMENTS	112.02	47.46
INCORRECT DOCUMENTS	97.80	42.20

TABLE 2 EFFECTIVENESS VS DOCUMENT LENGTH
35 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 7 48 DISCRIMINATING WORDS

TEST DOCUMENTS		CATEGORY P4			
NUMBER OF TOKENS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0 - 9	0.	0.	0.	0.	0.
10 - 19	0.	0.	0.	0.	0.
20 - 29	1.	1.	0.	1.00	0.
30 - 39	6.	6.	0.	1.00	0.
40 - 49	4.	2.	2.	0.50	0.50
50 - 59	13.	8.	5.	0.62	0.38
60 - 69	16.	12.	4.	0.75	0.25
70 - 79	11.	9.	2.	0.82	0.18
80 - 89	9.	6.	3.	0.67	0.33
90 - 99	9.	6.	3.	0.67	0.33
100 - 109	5.	2.	3.	0.40	0.60
110 - 119	6.	6.	0.	1.00	0.
120 - 129	5.	2.	3.	0.40	0.60
130 - 139	6.	3.	3.	0.50	0.50
140 - 149	9.	5.	4.	0.56	0.44
150 - 159	2.	0.	2.	0.	1.00
160 - 169	4.	3.	1.	0.75	0.25
170 - 179	1.	1.	0.	1.00	0.
180 - 189	1.	0.	1.	0.	1.00
190 - 199	0.	0.	0.	0.	0.
200 - 209	1.	0.	1.	0.	1.00
210 - 219	1.	1.	0.	1.00	0.
220 - 229	0.	0.	0.	0.	0.
230 - 239	1.	1.	0.	1.00	0.
240 - 249	0.	0.	0.	0.	0.
250 - UP	0.	0.	0.	0.	0.
TOTAL	111.	74.	37.	0.67	0.33

	MEAN	S.D.
CORRECT DOCUMENTS	89.42	43.22
INCORRECT DOCUMENTS	103.92	41.36

TABLE 2 EFFECTIVENESS VS DOCUMENT LENGTH
35 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 7 48 DISCRIMINATING WORDS

TEST DOCUMENTS			CATEGORY TOTAL		
NUMBER OF TOKENS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0 - 9	0.	0.	0.	0.	0.
10 - 19	0.	0.	0.	0.	0.
20 - 29	4.	2.	2.	0.50	0.50
30 - 39	24.	16.	8.	0.67	0.33
40 - 49	30.	22.	8.	0.73	0.27
50 - 59	49.	30.	19.	0.61	0.39
60 - 69	64.	48.	16.	0.75	0.25
70 - 79	58.	46.	12.	0.79	0.21
80 - 89	49.	34.	15.	0.69	0.31
90 - 99	51.	34.	17.	0.67	0.33
100 - 109	42.	31.	11.	0.74	0.26
110 - 119	42.	35.	7.	0.83	0.17
120 - 129	52.	43.	9.	0.83	0.17
130 - 139	35.	22.	13.	0.63	0.37
140 - 149	27.	19.	8.	0.70	0.30
150 - 159	20.	16.	4.	0.80	0.20
160 - 169	22.	16.	6.	0.73	0.27
170 - 179	22.	17.	5.	0.77	0.23
180 - 189	15.	13.	2.	0.87	0.13
190 - 199	8.	6.	2.	0.75	0.25
200 - 209	5.	4.	1.	0.80	0.20
210 - 219	3.	2.	1.	0.67	0.33
220 - 229	1.	1.	0.	1.00	0.
230 - 239	7.	7.	0.	1.00	0.
240 - 249	1.	1.	0.	1.00	0.
250 - UP	1.	1.	0.	1.00	0.
TOTAL	632.	466.	166.	0.74	0.26

	MEAN	S.D.
CORRECT DOCUMENTS	105.86	46.47
INCORRECT DOCUMENTS	97.30	42.37

TABLE 3 EFFECTIVENESS VS WORDS IN INTERSECTION OF DOCUMENTS AND DISCRIMINATING SET
35 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 7 48 DISCRIMINATING WORDS

TEST DOCUMENTS			CATEGORY P2		
NUMBER OF WORDS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0	5.	0.	5.	0.	1.00
1	24.	11.	13.	0.46	0.54
2	36.	29.	7.	0.81	0.19
3	50.	41.	9.	0.82	0.18
4	34.	25.	9.	0.74	0.26
5	38.	29.	9.	0.76	0.24
6	28.	19.	9.	0.68	0.32
7	38.	28.	10.	0.74	0.26
8	19.	14.	5.	0.74	0.26
9	9.	5.	4.	0.56	0.44
10	7.	4.	3.	0.57	0.43
11	0.	0.	0.	0.	0.
12	2.	2.	0.	1.00	0.
13	0.	0.	0.	0.	0.
14	0.	0.	0.	0.	0.
15	0.	0.	0.	0.	0.
16	0.	0.	0.	0.	0.
17	0.	0.	0.	0.	0.
18	0.	0.	0.	0.	0.
19	0.	0.	0.	0.	0.
20	0.	0.	0.	0.	0.
TOTAL	290.	207.	83.	0.71	0.29

	MEAN	S.D.
CORRECT DOCUMENTS	4.68	2.35
INCORRECT DOCUMENTS	4.40	2.79

TABLE 3 EFFECTIVENESS VS WORDS IN INTERSECTION OF DOCUMENTS AND DISCRIMINATING SET
35 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 7 48 DISCRIMINATING WORDS

TEST DOCUMENTS			CATEGORY P3		
NUMBER OF WORDS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0	1.	0.	1.	0.	1.00
1	4.	3.	1.	0.75	0.25
2	18.	11.	7.	0.61	0.39
3	27.	19.	8.	0.70	0.30
4	48.	42.	6.	0.88	0.13
5	34.	27.	7.	0.79	0.21
6	26.	20.	6.	0.77	0.23
7	28.	23.	5.	0.82	0.18
8	17.	14.	3.	0.82	0.18
9	14.	14.	0.	1.00	0.
10	10.	8.	2.	0.80	0.20
11	3.	3.	0.	1.00	0.
12	0.	0.	0.	0.	0.
13	0.	0.	0.	0.	0.
14	1.	1.	0.	1.00	0.
15	0.	0.	0.	0.	0.
16	0.	0.	0.	0.	0.
17	0.	0.	0.	0.	0.
18	0.	0.	0.	0.	0.
19	0.	0.	0.	0.	0.
20	0.	0.	0.	0.	0.
TOTAL	231.	185.	46.	0.80	0.20

	MEAN	S.D.
CORRECT DOCUMENTS	5.57	2.39
INCORRECT DOCUMENTS	4.63	2.28

TABLE 3 EFFECTIVENESS VS WORDS IN INTERSECTION OF DOCUMENTS AND DISCRIMINATING SET
25 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 7 48 DISCRIMINATING WORDS

	TEST DOCUMENTS		CATEGORY P4		
NUMBER OF WORDS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0	0.	0.	0.	0.	0.
1	6.	4.	2.	0.67	0.33
2	16.	11.	5.	0.69	0.31
3	20.	13.	7.	0.65	0.35
4	12.	11.	1.	0.92	0.08
5	17.	13.	4.	0.76	0.24
6	13.	8.	5.	0.62	0.38
7	10.	7.	3.	0.70	0.30
8	8.	1.	7.	0.13	0.88
9	5.	3.	2.	0.60	0.40
10	2.	1.	1.	0.50	0.50
11	1.	1.	0.	1.00	0.
12	1.	1.	0.	1.00	0.
13	0.	0.	0.	0.	0.
14	0.	0.	0.	0.	0.
15	0.	0.	0.	0.	0.
16	0.	0.	0.	0.	0.
17	0.	0.	0.	0.	0.
18	0.	0.	0.	0.	0.
19	0.	0.	0.	0.	0.
20	0.	0.	0.	0.	0.
TOTAL	111.	74.	37.	0.67	0.33

	MEAN	S.D.
CORRECT DOCUMENTS	4.58	2.37
INCORRECT DOCUMENTS	5.19	2.56

TABLE 3   EFFECTIVENESS VS WORDS IN INTERSECTION OF DOCUMENTS AND DISCRIMINATING SET
35 DOCUMENTS IN EACH CATEGORY   DISCRIMINATING WORD SET   7   48 DISCRIMINATING WORDS

TEST DOCUMENTS			CATEGORY TOTAL		
NUMBER OF WORDS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0	6.	0.	6.	0.	1.00
1	34.	18.	16.	0.53	0.47
2	70.	51.	19.	0.73	0.27
3	97.	73.	24.	0.75	0.25
4	94.	78.	16.	0.83	0.17
5	89.	69.	20.	0.78	0.22
6	67.	47.	20.	0.70	0.30
7	76.	58.	18.	0.76	0.24
8	44.	29.	15.	0.66	0.34
9	28.	22.	6.	0.79	0.21
10	19.	13.	6.	0.68	0.32
11	4.	4.	0.	1.00	0.
12	3.	3.	0.	1.00	0.
13	0.	0.	0.	0.	0.
14	1.	1.	0.	1.00	0.
15	0.	0.	0.	0.	0.
16	0.	0.	0.	0.	0.
17	0.	0.	0.	0.	0.
18	0.	0.	0.	0.	0.
19	0.	0.	0.	0.	0.
20	0.	0.	0.	0.	0.
TOTAL	632.	466.	166.	0.74	0.26

	MEAN	S.D.
CORRECT DOCUMENTS	5.02	2.41
INCORRECT DOCUMENTS	4.64	2.62

TABLE 4 EFFECTIVENESS VS RADIUS
35 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 7 48 DISCRIMINATING WORDS

TEST DOCUMENTS		CATEGORY P2			
RADIUS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0. - 0.49	145.	95.	47.	0.68	0.32
0.5 - 0.99	96.	83.	13.	0.86	0.14
1.0 - 1.99	38.	25.	13.	0.66	0.34
2.0 - 2.99	6.	1.	5.	0.17	0.83
3.0 - 3.99	3.	0.	3.	0.	1.00
4.0 - 4.99	2.	0.	2.	0.	1.00
5.0 - 5.99	0.	0.	0.	0.	0.
6.0 - 6.99	0.	0.	0.	0.	0.
7.0 - 7.99	0.	0.	0.	0.	0.
8.0 - 8.99	0.	0.	0.	0.	0.
9.0 - 9.99	0.	0.	0.	0.	0.
10. - UP	0.	0.	0.	0.	0.
TOTAL	290.	207.	83.	0.71	0.29

	MEAN	S.D.
CORRECT DOCUMENTS	0.59	0.37
INCORRECT DOCUMENTS	0.83	1.02

TABLE 4 EFFECTIVENESS VS RADIUS
35 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 7 48 DISCRIMINATING WORDS

	TEST DOCUMENTS		CATEGORY P3		
RADIUS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0. - 0.49	86.	51.	35.	0.59	0.41
0.5 - 0.99	94.	84.	10.	0.89	0.11
1.0 - 1.99	48.	47.	1.	0.98	0.02
2.0 - 2.99	3.	3.	0.	1.00	0.
3.0 - 3.99	0.	0.	0.	0.	0.
4.0 - 4.99	0.	0.	0.	0.	0.
5.0 - 5.99	0.	0.	0.	0.	0.
6.0 - 6.99	0.	0.	0.	0.	0.
7.0 - 7.99	0.	0.	0.	0.	0.
8.0 - 8.99	0.	0.	0.	0.	0.
9.0 - 9.99	0.	0.	0.	0.	0.
10. - UP	0.	0.	0.	0.	0.
TOTAL	231.	185.	46.	0.80	0.20

	MEAN	S.D.
CORRECT DOCUMENTS	0.79	0.43
INCORRECT DOCUMENTS	0.30	0.25

TABLE 4 EFFECTIVENESS VS RADIUS
35 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 7 48 DISCRIMINATING WORDS

TEST DOCUMENTS		CATEGORY P4			
RADIUS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0. - 0.49	71.	48.	23.	0.68	0.32
0.5 - 0.99	29.	18.	11.	0.62	0.38
1.0 - 1.99	10.	8.	2.	0.80	0.20
2.0 - 2.99	1.	0.	1.	0.	1.00
3.0 - 3.99	0.	0.	0.	0.	0.
4.0 - 4.99	0.	0.	0.	0.	0.
5.0 - 5.99	0.	0.	0.	0.	0.
6.0 - 6.99	0.	0.	0.	0.	0.
7.0 - 7.99	0.	0.	0.	0.	0.
8.0 - 8.99	0.	0.	0.	0.	0.
9.0 - 9.99	0.	0.	0.	0.	0.
10. - UP	0.	0.	0.	0.	0.
TOTAL	111.	74.	37.	0.67	0.33

	MEAN	S.D.
CORRECT DOCUMENTS	0.47	0.39
INCORRECT DOCUMENTS	0.48	0.43

TABLE 4 EFFECTIVENESS VS RADIUS
35 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 7 48 DISCRIMINATING WORDS

TEST DOCUMENTS			CATEGORY TOTAL		
RADIUS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0. - 0.49	302.	197.	105.	0.65	0.35
0.5 - 0.99	219.	185.	34.	0.84	0.16
1.0 - 1.99	96.	80.	16.	0.83	0.17
2.0 - 2.99	10.	4.	6.	0.40	0.60
3.0 - 3.99	3.	0.	3.	0.	1.00
4.0 - 4.99	2.	0.	2.	0.	1.00
5.0 - 5.99	0.	0.	0.	0.	0.
6.0 - 6.99	0.	0.	0.	0.	0.
7.0 - 7.99	0.	0.	0.	0.	0.
8.0 - 8.99	0.	0.	0.	0.	0.
9.0 - 9.99	0.	0.	0.	0.	0.
10. - UP	0.	0.	0.	0.	0.
TOTAL	632.	466.	166.	0.74	0.26

	MEAN	S.D.
CORRECT DOCUMENTS	0.65	0.42
INCORRECT DOCUMENTS	0.61	0.79

```
TABLE 5   DOCUMENT CLASSIFICATION SUMMARY
          35 DOCUMENTS IN EACH CATEGORY   DISCRIMINATING WORD SET    7      48 DISCRIMINATING  WORDS

  TEST
                 AUTO    CATEGORY
ACTUAL
CATEGORY         P2       P3       P4       TOTAL

P2             207.00    28.00    55.00   290.00

P3              29.00   185.00    17.00   231.00

P4              29.00     8.00    74.00   111.00

TOTAL          265.00   221.00   146.00   632.00

                   PERCENTAGE

P2               0.71     0.10     0.19     1.00

P3               0.13     0.80     0.07     1.00

P4               0.26     0.07     0.67     1.00

                           SWETS MEASURES            RECALL        RELEVANCE       PRECISION
CATEGORY P2            PERTINENT   NOT PERTINENT     RATIO           RATIO           RATIO
 RETRIEVED               0.71          0.13           0.71           0.78            0.91


                           SWETS MEASURES            RECALL        RELEVANCE       PRECISION
CATEGORY P3            PERTINENT   NOT PERTINENT     RATIO           RATIO           RATIO
 RETRIEVED               0.80          0.07           0.80           0.84            0.96


                           SWETS MEASURES            RECALL        RELEVANCE       PRECISION
CATEGORY P4            PERTINENT   NOT PERTINENT     RATIO           RATIO           RATIO
 RETRIEVED               0.67          0.11           0.67           0.51            1.32
```

TABLE 1 EFFECTIVENESS VS NUMBER OF SENTENCES
70 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 7 48 DISCRIMINATING WORDS

SAMPLE DOCUMENTS			CATEGORY P2		
NUMBER OF SENTENCES	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
1	1.	1.	0.	1.00	0.
2	4.	3.	1.	0.75	0.25
3	8.	7.	1.	0.88	0.13
4	15.	14.	1.	0.93	0.07
5	10.	10.	0.	1.00	0.
6	16.	16.	0.	1.00	0.
7	6.	6.	0.	1.00	0.
8	3.	3.	0.	1.00	0.
9	1.	1.	0.	1.00	0.
10	2.	2.	0.	1.00	0.
11	2.	2.	0.	1.00	0.
12	2.	2.	0.	1.00	0.
13	0.	0.	0.	0.	0.
14	0.	0.	0.	0.	0.
15	0.	0.	0.	0.	0.
TOTAL	70.	67.	3.	0.96	0.04

	MEAN	S.D.
CORRECT DOCUMENTS	5.54	2.35
INCORRECT DOCUMENTS	3.00	0.82

TABLE 1 EFFECTIVENESS VS NUMBER OF SENTENCES
70 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 7 48 DISCRIMINATING WORDS

SAMPLE DOCUMENTS			CATEGORY P3		
NUMBER OF SENTENCES	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
1	1.	0.	1.	0.	1.00
2	6.	4.	2.	0.67	0.33
3	12.	12.	0.	1.00	0.
4	10.	10.	0.	1.00	0.
5	9.	9.	0.	1.00	0.
6	10.	9.	1.	0.90	0.10
7	9.	9.	0.	1.00	0.
8	3.	3.	0.	1.00	0.
9	5.	5.	0.	1.00	0.
10	3.	3.	0.	1.00	0.
11	1.	1.	0.	1.00	0.
12	1.	1.	0.	1.00	0.
13	0.	0	0.	0.	0.
14	0.	0.	0.	0.	0.
15	0.	0.	0.	0.	0.
TOTAL	70.	66.	4.	0.94	0.06

	MEAN	S.D.
CORRECT DOCUMENTS	5.58	2.41
INCORRECT DOCUMENTS	2.75	1.92

TABLE 1 EFFECTIVENESS VS NUMBER OF SENTENCES
70 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 7 40 DISCRIMINATING WORDS

SAMPLE DOCUMENTS NUMBER OF SENTENCES	 NUMBER OF DOCUMENTS	 NUMBER OF CORRECT DOCUMENTS	CATEGORY PO NUMBER OF INCORRECT DOCUMENTS	 PERCENTAGE OF CORRECT DOCUMENTS	 PERCENTAGE OF INCORRECT DOCUMENTS
1	1.	1.	0.	1.00	0.
2	10.	9.	1.	0.90	0.10
3	12.	10.	2.	0.83	0.17
4	13.	12.	1.	0.92	0.08
5	7.	6.	1.	0.86	0.14
6	8.	8.	0.	1.00	0.
7	9.	8.	1.	0.89	0.11
8	4.	4.	0.	1.00	0.
9	1.	1.	0.	1.00	0.
10	4.	4.	0.	1.00	0.
11	1.	1.	0.	1.00	0.
12	0.	0.	0.	0.	0.
13	0.	0.	0.	0.	0.
14	0.	0.	0.	0.	0.
15	0.	0.	0.	0.	0.
TOTAL	70.	64.	6.	0.91	0.09

	MEAN	S.D.
CORRECT DOCUMENTS	5.05	2.43
INCORRECT DOCUMENTS	4.00	1.63

TABLE 1 EFFECTIVENESS VS NUMBER OF SEN ENCES
70 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 7 48 DISCRIMINATING WORDS

SAMPLE DOCUMENTS			CATEGORY TOTAL		
NUMBER OF SENTENCES	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
1	3.	2.	1.	0.67	0.33
2	20.	16.	4.	0.80	0.20
3	32.	29.	3.	0.91	0.09
4	38.	36.	2.	0.95	0.05
5	26.	25.	1.	0.96	0.04
6	34.	33.	1.	0.97	0.03
7	24.	23.	1.	0.96	0.04
8	10.	10.	0.	1.00	0.
9	7.	7.	0.	1.00	0.
10	9.	9.	0.	1.00	0.
11	4.	4.	0.	1.00	0.
12	3.	3.	0.	1.00	0.
13	0.	0.	0.	0.	0.
14	0.	0.	0.	0.	0.
15	0.	0.	0.	0.	0.
TOTAL	210.	197.	13.	0.94	0.06

	MEAN	S.D.
CORRECT DOCUMENTS	5.39	2.41
INCORRECT DOCUMENTS	3.38	1.69

```
TABLE 2   EFFECTIVENESS VS DOCUMENT LENGTH
          70 DOCUMENTS IN EACH CATEGORY   DISCRIMINATING WORD SET    7     48 DISCRIMINATING  WORDS

  SAMPLE DOCUMENTS                            CATEGORY P2
NUMBER OF     NUMBER OF   NUMBER OF             NUMBER OF              PERCENTAGE OF        PERCENTAGE OF
 TOKENS       DOCUMENTS    CORRECT DOCUMENTS   INCORRECT DOCUMENTS     CORRECT DOCUMENTS    INCORRECT DOCUMENTS
   0 -     9    0.               0.                 0.                   0.                   0.
  10 -    19    0.               0.                 0.                   0.                   0.
  20 -    29    0.               0.                 0.                   0.                   0.
  30 -    39    3.               2.                 1.                   0.67                 0.33
  40 -    49    5.               5.                 0.                   1.00                 0.
  50 -    59    4.               4.                 0.                   1.00                 0.
  60 -    69    5.               5.                 0.                   1.00                 0.
  70 -    79    7.               7.                 0.                   1.00                 0.
  80 -    89    7.               6.                 1.                   0.86                 0.14
  90 -    99    6.               5.                 1.                   0.83                 0.17
 100 -   109    3.               3.                 0.                   1.00                 0.
 110 -   119    6.               6.                 0.                   1.00                 0.
 120 -   129    6.               6.                 0.                   1.00                 0.
 130 -   139    4.               4.                 0.                   1.00                 0.
 140 -   149    5.               5.                 0.                   1.00                 0.
 150 -   159    1.               1.                 0.                   1.00                 0.
 160 -   169    2.               2.                 0.                   1.00                 0.
 170 -   179    1.               1.                 0.                   1.00                 0.
 180 -   189    1.               1.                 0.                   1.00                 0.
 190 -   199    2.               2.                 0.                   1.00                 0.
 200 -   209    0.               0.                 0.                   0.                   0.
 210 -   219    0.               0.                 0.                   0.                   0.
 220 -   229    0.               0.                 0.                   0.                   0.
 230 -   239    1.               1.                 0.                   1.00                 0.
 240 -   249    0.               0.                 0.                   0.                   0.
 250 -    UP    1.               1.                 0.                   1.00                 0.

TOTAL          70.              67.                 3.                   0.96                 0.04

                          MEAN              S.D.
CORRECT    DOCUMENTS     105.51             47.21
INCORRECT DOCUMENTS       68.00             26.42
```

TABLE 2 EFFECTIVENESS VS DOCUMENT LENGTH

70 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 7 50 DISCRIMINATING WORDS

SAMPLE DOCUMENTS		CATEGORY P3			
NUMBER OF TOKENS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0 - 9	0.	0.	0.	0.	0.
10 - 19	0.	0.	0.	0.	0.
20 - 29	2.	1.	1.	0.50	0.50
30 - 39	2.	1.	1.	0.50	0.50
40 - 49	3.	3.	0.	1.00	0.
50 - 59	2.	1.	1.	0.50	0.50
60 - 69	5.	5.	0.	1.00	0.
70 - 79	7.	7.	0.	1.00	0.
80 - 89	3.	3.	0.	1.00	0.
90 - 99	11.	10.	1.	0.91	0.09
100 - 109	2.	2.	0.	1.00	0.
110 - 119	7.	7.	0.	1.00	0.
120 - 129	6.	6.	0.	1.00	0.
130 - 139	1.	1.	0.	1.00	0.
140 - 149	5.	5.	0.	1.00	0.
150 - 159	1.	1.	0.	1.00	0.
160 - 169	0.	0.	0.	0.	0.
170 - 179	1.	1.	0.	1.00	0.
180 - 189	4.	4.	0.	1.00	0.
190 - 199	3.	3.	0.	1.00	0.
200 - 209	2.	2.	0.	1.00	0.
210 - 219	1.	1.	0.	1.00	0.
220 - 229	0.	0.	0.	0.	0.
230 - 239	1.	1.	0.	1.00	0.
240 - 249	1.	1.	0.	1.00	0.
250 - UP	0.	0.	0.	0.	0.
TOTAL	70.	66.	4.	0.94	0.06

	MEAN	S.D.
CORRECT DOCUMENTS	115.85	50.27
INCORRECT DOCUMENTS	52.25	28.09

TABLE 2 EFFECTIVENESS VS DOCUMENT LENGTH
70 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 7 48 DISCRIMINATING WORDS

SAMPLE DOCUMENTS		CATEGORY P4			
NUMBER OF TOKENS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0 - 9	0.	0.	0.	0.	0.
10 - 19	0.	0.	0.	0.	0.
20 - 29	1.	0.	1.	0.	1.00
30 - 39	3.	3.	0.	1.00	0.
40 - 49	3.	3.	0.	1.00	0.
50 - 59	7.	6.	1.	0.86	0.14
60 - 69	10.	9.	1.	0.90	0.10
70 - 79	11.	10.	1.	0.91	0.09
80 - 89	3.	3.	0.	1.00	0.
90 - 99	3.	3.	0.	1.00	0.
100 - 109	4.	3.	1.	0.75	0.25
110 - 119	5.	5.	0.	1.00	0.
120 - 129	4.	4.	0.	1.00	0.
130 - 139	6.	5.	1.	0.83	0.17
140 - 149	5.	5.	0.	1.00	0.
150 - 159	3.	3.	0.	1.00	0.
160 - 169	1.	1.	0.	1.00	0.
170 - 179	0.	0.	0.	0.	0.
180 - 189	1.	1.	0.	1.00	0.
190 - 199	0.	0.	0.	0.	0.
200 - 209	0.	0.	0.	0.	0.
210 - 219	0.	0.	0.	0.	0.
220 - 229	C.	0.	0.	0.	0.
230 - 239	0.	0.	0.	0.	0.
240 - 249	0.	0.	0.	0.	0.
250 - UP	0.	0.	0.	0.	0.
TOTAL	70.	64.	6.	0.91	0.09

	MEAN	S.D.
CORRECT DOCUMENTS	93.73	37.95
INCORRECT DOCUMENTS	76.00	34.42

TABLE 2 EFFECTIVENESS VS DOCUMENT LENGTH
70 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 7 49 DISCRIMINATING WORDS

SAMPLE DOCUMENTS			CATEGORY TOTAL		
NUMBER OF TOKENS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0 - 9	0.	0.	0.	0.	0.
10 - 19	0.	0.	0.	0.	0.
20 - 29	3.	1.	2.	0.33	0.67
30 - 39	8.	6.	2.	0.75	0.25
40 - 49	11.	11.	0.	1.00	0.
50 - 59	13.	11.	2.	0.85	0.15
60 - 69	20.	19.	1.	0.95	0.05
70 - 79	25.	24.	1.	0.96	0.04
80 - 89	13.	12.	1.	0.92	0.08
90 - 99	20.	18.	2.	0.90	0.10
100 - 109	9.	8.	1.	0.89	0.11
110 - 119	18.	18.	0.	1.00	0.
120 - 129	16.	16.	0.	1.00	0.
130 - 139	11.	10.	1.	0.91	0.09
140 - 149	15.	15.	0.	1.00	0.
150 - 159	5.	5.	0.	1.00	0.
160 - 169	3.	3.	0.	1.00	0.
170 - 179	2.	2.	0.	1.00	0.
180 - 189	6.	6.	0.	1.00	0.
190 - 199	5.	5.	0.	1.00	0.
200 - 209	2.	2.	0.	1.00	0.
210 - 219	1.	1.	0.	1.00	0.
220 - 229	0.	0.	0.	0.	0.
230 - 239	2.	2.	0.	1.00	0.
240 - 249	1.	1.	0.	1.00	0.
250 - UP	1.	1.	0.	1.00	0.
TOTAL	210.	197.	13.	0.94	0.06

	MEAN	S.D.
CORRECT DOCUMENTS	105.15	46.40
INCORRECT DOCUMENTS	66.85	32.49

TABLE 3 EFFECTIVENESS VS WORDS IN INTERSECTION OF DOCUMENTS AND DISCRIMINATING SET
70 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 7 48 DISCRIMINATING WORDS

SAMPLE	DOCUMENTS		CATEGORY P2		
NUMBER OF WORDS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0	0.	0.	0.	0.	0.
1	4.	3.	1.	0.75	0.25
2	8.	8.	0.	1.00	0.
3	12.	12.	0.	1.00	0.
4	6.	6.	0.	1.00	0.
5	12.	11.	1.	0.92	0.08
6	10.	10.	0.	1.00	0.
7	11.	10.	1.	0.91	0.09
8	4.	4.	0.	1.00	0.
9	2.	2.	0.	1.00	0.
10	1.	1.	0.	1.00	0.
11	0.	0.	0.	0.	0.
12	0.	0.	0.	0.	0.
13	0.	0.	0.	0.	0.
14	0.	0.	0.	0.	0.
15	0.	0.	0.	0.	0.
16	0.	0.	0.	0.	0.
17	0.	0.	0.	0.	0.
18	0.	0.	0.	0.	0.
19	0.	0.	0.	0.	0.
20	0.	0.	0.	0.	0.
TOTAL	70.	67.	3.	0.96	0.04

	MEAN	S.D.
CORRECT DOCUMENTS	4.84	2.17
INCORRECT DOCUMENTS	4.33	2.49

TABLE 3 EFFECTIVENESS VS WORDS IN INTERSECTION OF DOCUMENTS AND DISCRIMINATING SET
70 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 7 48 DISCRIMINATING WORDS

SAMPLE DOCUMENTS		CATEGORY P3			
NUMBER OF WORDS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0	0.	0.	0.	0.	0.
1	1.	0.	1.	0.	1.00
2	6.	6.	0.	1.00	0.
3	6.	5.	1.	0.83	0.17
4	12.	12.	0.	1.00	0.
5	9.	8.	1.	0.89	0.11
6	10.	10.	0.	1.00	0.
7	9.	8.	1.	0.89	0.11
8	8.	8.	0.	1.00	0.
9	6.	6.	0.	1.00	0.
10	2.	2.	0.	1.00	0.
11	0.	0.	0.	0.	0.
12	0.	0.	0.	0.	0.
13	0.	0.	0.	0.	0.
14	1.	1.	0.	1.00	0.
15	0.	0.	0.	0.	0.
16	0.	0.	0.	0.	0.
17	0.	0.	0.	0.	0.
18	0.	0.	0.	0.	0.
19	0.	0.	0.	0.	0.
20	0.	0.	0.	0.	0.
TOTAL	70.	66.	4.	0.94	0.06

	MEAN	S.D.
CORRECT DOCUMENTS	5.80	2.41
INCORRECT DOCUMENTS	4.00	2.24

TABLE 3 EFFECTIVENESS VS WORDS IN INTERSECTION OF DOCUMENTS AND DISCRIMINATING SET
70 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 7 48 DISCRIMINATING WORDS

SAMPLE DOCUMENTS		CATEGORY P4			
NUMBER OF WORDS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0	0.	0.	0.	0.	0.
1	3.	2.	1.	0.67	0.33
2	9.	7.	2.	0.78	0.22
3	10.	10.	0.	1.00	0.
4	7.	7.	0.	1.00	0.
5	12.	11.	1.	0.92	0.08
6	9.	9.	0.	1.00	0.
7	7.	7.	0.	1.00	0.
8	5.	4.	1.	0.80	0.20
9	6.	6.	0.	1.00	0.
10	2.	1.	1.	0.50	0.50
11	0.	0.	0.	0.	0.
12	0.	0.	0.	0.	0.
13	0.	0.	0.	0.	0.
14	0.	0.	0.	0.	0.
15	0.	0.	0.	0.	0.
16	0.	0.	0.	0.	0.
17	0.	0.	0.	0.	0.
18	0.	0.	0.	0.	0.
19	0.	0.	0.	0.	0.
20	0.	0.	0.	0.	0.
TOTAL	70.	64.	6.	0.91	0.09

	MEAN	S.D.
CORRECT DOCUMENTS	5.13	2.29
INCORRECT DOCUMENTS	4.67	3.35

TABLE 3 EFFECTIVENESS VS WORDS IN INTERSECTION OF DOCUMENTS AND DISCRIMINATING SET
70 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 7 48 DISCRIMINATING WORDS

SAMPLE DOCUMENTS			CATEGORY TOTAL		
NUMBER OF WORDS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0	0.	0.	0.	0.	0.
	8.	5.	3.	0.63	0.38
2	23.	21.	2.	0.91	0.09
3	28.	27.	1.	0.96	0.04
4	25.	25.	0.	1.00	0.
5	33.	30.	3.	0.91	0.09
6	29.	29.	0.	1.00	0.
7	27.	25.	2.	0.93	0.07
8	17.	16.	1.	0.94	0.06
9	14.	14.	0.	1.00	0.
10	5.	4.	1.	0.80	0.20
11	0.	0.	0.	0.	0.
12	0.	0.	0.	0.	0.
13	0.	0.	0.	0.	0.
14	1.	1.	0.	1.00	0.
15	0.	0.	0.	0.	0.
16	0.	0.	0.	0.	0.
17	0.	0.	0.	0.	0.
18	0.	0.	0.	0.	0.
19	0.	0.	0.	0.	0.
20	0.	0.	0.	0.	0.
TOTAL	210.	197.	13.	0.94	0.06

	MEAN	S.D.
CORRECT DOCUMENTS	5.25	2.33
INCORRECT DOCUMENTS	4.38	2.87

TABLE 4 EFFECTIVENESS VS RADIUS
70 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 7 48 DISCRIMINATING WORDS

SAMPLE DOCUMENTS			CATEGORY P2		
RADIUS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0. - 0.49	25.	23.	2.	0.92	0.08
0.5 - 0.99	28.	27.	1.	0.96	0.04
1.0 - 1.99	17.	17.	0.	1.00	0.
2.0 - 2.99	0.	0.	0.	0.	0.
3.0 - 3.99	0.	0.	0.	0.	0.
4.0 - 4.99	0.	0.	0.	0.	0.
5.0 - 5.99	0.	0.	0.	0.	0.
6.0 - 6.99	0.	0.	0.	0.	0.
7.0 - 7.99	0.	0.	0.	0.	0.
8.0 - 0.99	0.	0.	0.	0.	0.
9.0 - 9.99	0.	0.	0.	0.	0.
10. - UP	0.	0.	0.	0.	0.
TOTAL	70.	67.	3.	0.96	0.04

	MEAN	S.D.
CORRECT DOCUMENTS	0.72	0.44
INCORRECT DOCUMENTS	0.37	0.26

TABLE 4 EFFECTIVENESS VS RADIUS
70 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 7 48 DISCRIMINATING WORDS

SAMPLE DOCUMENTS			CATEGORY P3		
RADIUS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0. - 0.49	10.	6.	4.	0.60	0.40
0.5 - 0.99	29.	29.	0.	1.00	0.
1.0 - 1.99	31.	31.	0.	1.00	0.
2.0 - 2.99	0.	0.	0.	0.	0.
3.0 - 3.99	0.	0.	0.	0.	0.
4.0 - 4.99	0.	0.	0.	0.	0.
5.0 - 5.99	0.	0.	0.	0.	0.
6.0 - 6.99	0.	0.	0.	0.	0.
7.0 - 7.99	0.	0.	0.	0.	0.
8.0 - 8.99	0.	0.	0.	0.	0.
9.0 - 9.99	0.	0.	0.	0.	0.
10. - UP	0.	0.	0.	0.	0.
TOTAL	70.	66.	4.	0.94	0.06

	MEAN	S.D.
CORRECT DOCUMENTS	0.9	0.34
INCORRECT DOCUMENTS	0.22	0.10

TABLE 4 EFFECTIVENESS VS RADIUS
70 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 7 48 DISCRIMINATING WORDS

SAMPLE DOCUMENTS		CATEGORY P4			
RADIUS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0. - 0.49	15.	14.	4.	0.78	0.22
0.5 - 0.99	26.	25.	1.	0.96	0.04
1.0 - 1.99	24.	23.	1.	0.96	0.04
2.0 - 2.99	2.	2.	0.	1.00	0.
3.0 - 3.99	0.	0.	0.	0.	0.
4.0 - 4.99	0.	0.	0.	0.	0.
5.0 - 5.99	0.	0.	0.	0.	0.
6.0 - 6.99	0.	0.	0.	0.	0.
7.0 - 7.99	0.	0.	0.	0.	0.
8.0 - 8.99	0.	0.	0.	0.	0.
9.0 - 9.99	0.	0.	0.	0.	0.
10. - UP	0.	0.	0.	0.	0.
TOTAL	70.	64.	6.	0.91	0.09

	MEAN	S.D.
CORRECT DOCUMENTS	0.89	0.46
INCORRECT DOCUMENTS	0.37	0.44

TABLE 4 EFFECTIVENESS VS RADIUS
70 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 7 48 DISCRIMINATING WORDS

SAMPLE DOCUMENTS			CATEGORY TOTAL		
RADIUS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0. - 0.49	53.	43.	10.	0.81	0.19
0.5 - 0.99	83.	81.	2.	0.98	0.02
1.0 - 1.99	72.	71.	1.	0.99	0.01
2.0 - 2.99	2.	2.	0.	1.00	0.
3.0 - 3.99	0.	0.	0.	0.	0.
4.0 - 4.99	0.	0.	0.	0.	0.
5.0 - 5.99	0.	0.	0.	0.	0.
6.0 - 6.99	0.	0.	0.	0.	0.
7.0 - 7.99	0.	0.	0.	0.	0.
8.0 - 8.99	0.	0.	0.	0.	0.
9.0 - 9.99	0.	0.	0.	0.	0.
10. - UP	0.	0.	0.	0.	0.
TOTAL	210.	197.	13.	0.94	0.06

	MEAN	S.D.
CORRECT DOCUMENTS	0.86	0.43
INCORRECT DOCUMENTS	0.33	0.34

```
TABLE 5   DOCUMENT CLASSIFICATION SUMMARY
          70 DOCUMENTS IN EACH CATEGORY   DISCRIMINATING WORD SET    7     48 DISCRIMINATING  WORDS

SAMPLE
                    AUTO   CATEGORY
ACTUAL
CATEGORY         P2        P3        P4       TOTAL

P2              67.00      1.00      2.00     70.00

P3               3.00     66.00      1.00     70.00

P4               5.00      1.00     64.00     70.00

TOTAL           75.00     68.00     67.00    210.00

                      PERCENTAGE

P2               0.96      0.01      0.03      1.00

P3               0.04      0.94      0.01      1.00

P4               0.07      0.01      0.91      1.00

                          SWETS MEASURES                 RECALL         RELEVANCE       PRECISION
CATEGORY P2          PERTINENT    NOT PERTINENT          RATIO           RATIO           RATIO
 RETRIEVED             0.96           0.06                0.96            0.89            1.07


                          SWETS MEASURES                 RECALL         RELEVANCE       PRECISION
CATEGORY P3          PERTINENT    NOT PERTINENT          RATIO           RATIO           RATIO
 RETRIEVED             0.94           0.01                0.94            0.97            0.97


                          SWETS MEASURES                 RECALL         RELEVANCE       PRECISION
CATEGORY P4          PERTINENT    NOT PERTINENT          RATIO           RATIO           RATIO
 RETRIEVED             0.91           0.02                0.91            0.96            0.96
```

TABLE 1 EFFECTIVENESS VS NUMDER OF SENTENCES
70 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 7 48 DISCRIMINATING WORDS

TEST DOCUMENTS			CATEGORY P2		
NUMBER OF SENTENCES	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
1	1.	1.	0.	1.00	0.
2	17.	14.	3.	0.82	0.18
3	45.	38.	7.	0.84	0.16
4	45.	36.	9.	0.80	0.20
5	49.	39.	10.	0.80	0.20
6	34.	24.	10.	0.71	0.29
7	22.	19.	3.	0.86	0.14
8	21.	17.	4.	0.81	0.19
9	13.	10.	3.	0.77	0.23
10	6.	5.	1.	0.83	0.17
11	2.	2.	0.	1.00	0.
12	0.	0.	0.	0.	0.
13	0.	0.	0.	0.	0.
14	0.	0.	0.	0.	0.
15	0.	0.	0.	0.	0.
TOTAL	255.	205.	50.	0.80	0.20

	MEAN	S.D.
CORRECT DOCUMENTS	5.16	2.14
INCORRECT DOCUMENTS	5.26	1.96

TABLE 1 EFFECTIVENESS VS NUMBER OF SENTENCES
70 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 7 48 DISCRIMINATING WORDS

TEST DOCUMENTS			CATEGORY P3		
NUMBER OF SENTENCES	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
1	3.	2.	1.	0.67	0.33
2	12.	9.	3.	0.75	0.25
3	19.	18.	1.	0.95	0.05
4	45.	40.	5.	0.89	0.11
5	34.	30.	4.	0.88	0.12
6	29.	28.	1.	0.97	0.03
7	21.	19.	2.	0.90	0.10
8	11.	9.	2.	0.82	0.18
9	13.	12.	1.	0.92	0.08
10	5.	4.	1.	0.80	0.20
11	2.	2.	0.	1.00	0.
12	1.	1.	0.	1.00	0.
13	0.	0.	0.	0.	0.
14	1.	1.	0.	1.00	0.
15	0.	0.	0.	0.	0.
TOTAL	196.	175.	21.	0.89	0.11

	MEAN	S.D.
CORRECT DOCUMENTS	5.45	2.23
INCORRECT DOCUMENTS	5.00	2.39

TABLE 1 EFFECTIVENESS VS NUMBER OF SENTENCES
70 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 7 48 DISCRIMINATING WORDS

TEST DOCUMENTS			CATEGORY P4		
NUMBER OF SENTENCES	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
1	1.	1.	0.	1.00	0.
2	6.	3.	3.	0.50	0.50
3	18.	15.	3.	0.83	0.17
4	10.	10.	0.	1.00	0.
5	15.	11.	4.	0.73	0.27
6	9.	9.	0.	1.00	0.
7	7.	7.	0.	1.00	0.
8	7.	5.	2.	0.71	0.29
9	1.	0.	1.	0.	1.00
10	0.	0.	0.	0.	0.
11	1.	1.	0.	1.00	0.
12	0.	0.	0.	0.	0.
13	0.	0.	0.	0.	0.
14	1.	1.	0.	1.00	0.
15	0.	0.	0.	0.	0.
TOTAL	76.	63.	13.	0.83	0.17

	MEAN	S.D.
CORRECT DOCUMENTS	5.00	2.23
INCORRECT DOCUMENTS	4.62	2.34

TABLE 1 EFFECTIVENESS VS NUMBER OF SENTENCES
70 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 7 48 DISCRIMINATING WORDS

TEST DOCUMENTS			CATEGORY TOTAL		
NUMBER OF SENTENCES	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
1	5.	4.	1.	0.80	0.20
2	35.	26.	9.	0.74	0.26
3	82.	71.	11.	0.87	0.13
4	100.	86.	14.	0.86	0.14
5	98.	80.	18.	0.82	0.18
6	72.	61.	11.	0.85	0.15
7	50.	45.	5.	0.90	0.10
8	39.	31.	8.	0.79	0.21
9	27.	22.	5.	0.81	0.19
10	11.	9.	2.	0.82	0.18
11	5.	5.	0.	1.00	0.
12	1.	1.	0.	1.00	0.
13	0.	0.	0.	0.	0.
14	2.	2.	0.	1.00	0.
15	0.	0.	0.	0.	0.
TOTAL	527.	443.	84.	0.84	0.16

	MEAN	S.D.
CORRECT DOCUMENTS	5.25	2.19
INCORRECT DOCUMENTS	5.10	2.15

TABLE 2 EFFECTIVENESS VS DOCUMENT LENGTH
70 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 7 18 DISCRIMINATING WORDS

TEST DOCUMENTS			CATEGORY P2		
NUMBER OF TOKENS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0 - 9	0.	0.	0.	0.	0.
10 - 19	0.	0.	0.	0.	0.
20 - 29	2.	2.	0.	1.00	0.
30 - 39	10.	7.	3.	0.70	0.30
40 - 49	12.	11.	1.	0.92	0.08
50 - 59	22.	16.	6.	0.73	0.27
60 - 69	28.	23.	5.	0.82	0.18
70 - 79	21.	17.	4.	0.81	0.19
80 - 89	22.	19.	3.	0.86	0.14
90 - 99	17.	14.	3.	0.82	0.18
100 - 109	17.	13.	4.	0.76	0.24
110 - 119	12.	12.	0.	1.00	0.
120 - 129	23.	20.	3.	0.87	0.13
130 - 139	14.	12.	2.	0.86	0.14
140 - 149	8.	5.	3.	0.63	0.38
150 - 159	13.	11.	2.	0.85	0.15
160 - 169	9.	7.	2.	0.78	0.22
170 - 179	12.	7.	5.	0.58	0.42
180 - 189	6.	3.	3.	0.50	0.50
190 - 199	2.	1.	1.	0.50	0.50
200 - 209	2.	2.	0.	1.00	0.
210 - 219	0.	0.	0.	0.	0.
220 - 229	1.	1.	0.	1.00	0.
230 - 239	2.	2.	0.	1.00	0.
240 - 249	0.	0.	0.	0.	0
250 - UP	0.	0.	0.	0.	0.
TOTAL	255.	205.	50.	0.80	0.20

	MEAN	S.D.
CORRECT DOCUMENTS	101.43	44.10
INCORRECT DOCUMENTS	107.70	48.32

TABLE 2 EFFECTIVENESS VS DOCUMENT LENGTH
70 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 7 48 DISCRIMINATING WORDS

TEST DOCUMENTS			CATEGORY P3		
NUMBER OF TOKENS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0 - 9	0.	0.	0.	0.	0.
10 - 19	0.	0.	0.	0.	0.
20 - 29	0.	0.	0.	0.	0.
30 - 39	7.	6.	1.	0.86	0.14
40 - 49	11.	10.	1.	0.91	0.09
50 - 59	14.	11.	3.	0.79	0.21
60 - 69	17.	15.	2.	0.88	0.12
70 - 79	16.	14.	2.	0.88	0.13
80 - 89	15.	13.	2.	0.87	0.13
90 - 99	12.	12.	0.	1.00	0.
100 - 109	19.	18.	1.	0.95	0.05
110 - 119	14.	11.	3.	0.79	0.21
120 - 129	16.	15.	1.	0.94	0.06
130 - 139	12	11.	1.	0.92	0.08
140 - 149	4.	4.	0.	1.00	0.
150 - 159	6.	5.	1.	0.83	0.17
160 - 169	10.	9.	1.	0.90	0.10
170 - 179	11.	10.	1.	0.91	0.09
180 - 189	5.	5.	0.	1.00	0.
190 - 199	3.	2.	1.	0.67	0.33
200 - 209	1.	1.	0.	1.00	0.
210 - 219	1.	1.	0.	1.00	0.
220 - 229	0.	0.	0.	0.	0.
230 - 239	2.	2.	0.	1.00	0.
240 - 249	0.	0.	0.	0.	0.
250 - UP	0.	0.	0.	0.	0.
TOTAL	196.	175.	21.	0.89	0.11

	MEAN	S.D.
CORRECT DOCUMENTS	107.31	44.83
INCORRECT DOCUMENTS	99.67	45.30

TABLE 2 EFFECTIVENESS VS DOCUMENT LENGTH
70 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 7 48 DISCRIMINATING WORDS

TEST DOCUMENTS			CATEGORY P4		
NUMBER OF TOKENS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0 - 9	0.	0.	0.	0.	0.
10 - 19	0.	0.	0.	0.	0.
20 - 29	0.	0.	0.	0.	0.
30 - 39	5.	3.	2.	0.60	0.40
40 - 49	3.	2.	1.	0.67	0.33
50 - 59	8.	7.	1.	0.88	0.13
60 - 69	11.	8.	3.	0.73	0.27
70 - 79	4.	4.	0.	1.00	0.
80 - 89	9.	9.	0.	1.00	0.
90 - 99	8.	7.	1.	0.88	0.13
100 - 109	6.	5.	1.	0.83	0.17
110 - 119	3.	3.	0.	1.00	0.
120 - 129	5.	4.	1.	0.80	0.20
130 - 139	1.	1.	0.	1.00	0.
140 - 149	5.	3.	2.	0.60	0.40
150 - 159	0.	0.	0.	0.	0.
160 - 169	4.	4.	0.	1.00	0.
170 - 179	1.	1.	0.	1.00	0.
180 - 189	0.	0.	0.	0.	0.
190 - 199	0.	0.	0.	0.	0.
200 - 209	1.	0.	1.	0.	1.00
210 - 219	1.	1.	0.	1.00	0.
220 - 229	0.	0.	0.	0.	0.
230 - 239	1.	1.	0.	1.00	0.
240 - 249	0.	0.	0.	0.	0.
250 - UP	0.	0.	0.	0.	0.
TOTAL	76.	63.	13.	0.83	0.17

	MEAN	S.D.
CORRECT DOCUMENTS	95.67	42.57
INCORRECT DOCUMENTS	89.77	49.92

TABLE 2 EFFECTIVENESS VS DOCUMENT LENGTH
70 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 7 43 DISCRIMINATING WORDS

TEST DOCUMENTS		CATEGORY TOTAL			
NUMBER OF TOKENS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0 - 9	0.	0.	0.	0.	0.
10 - 19	0.	0.	0.	0.	0.
20 - 29	2.	2.	0.	1.00	0.
30 - 39	22.	16.	6.	0.73	0.27
40 - 49	26.	23.	3.	0.88	0.12
50 - 59	44.	34.	10.	0.77	0.23
60 - 69	56.	46.	10.	0.82	0.18
70 - 79	41.	35.	6.	0.85	0.15
80 - 89	46.	41.	5.	0.89	0.11
90 - 99	37.	33.	4.	0.89	0.11
100 - 109	42.	36.	6.	0.86	0.14
110 - 119	29.	26.	3.	0.90	0.10
120 - 129	44.	39.	5.	0.89	0.11
130 - 139	27.	24.	3.	0.89	0.11
140 - 149	17.	12.	5.	0.71	0.29
150 - 159	19.	16.	3.	0.84	0.16
160 - 169	23.	20.	3.	0.87	0.13
170 - 179	24.	18.	6.	0.75	0.25
180 - 189	11.	8.	3.	0.73	0.27
190 - 199	5.	3.	2.	0.60	0.40
200 - 209	4.	3.	1.	0.75	0.25
210 - 219	2.	2.	0.	1.00	0.
220 - 229	1.	1.	0.	1.00	0.
230 - 239	5.	5.	0.	1.00	0.
240 - 249	0.	0.	0.	0.	0.
250 - UP	0.	0.	0.	0.	0.
TOTAL	527.	443.	84.	0.84	0.16

	MEAN	S.D.
CORRECT DOCUMENTS	102.93	44.36
INCORRECT DOCUMENTS	102.92	48.29

TABLE 3 EFFECTIVENESS VS WORDS IN INTERSECTION OF DOCUMENTS AND DISCRIMINATING SET
70 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 7 48 DISCRIMINATING WORDS

TEST DOCUMENTS			CATEGORY 92		
NUMBER OF WORDS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0	6.	6.	0.	1.00	0.
1	23.	22.	1.	0.96	0.04
2	32.	21.	11.	0.66	0.34
3	42.	34.	8.	0.81	0.19
4	34.	28.	6.	0.82	0.18
5	33.	27.	6.	0.82	0.18
6	23.	18.	5.	0.78	0.22
7	30.	25.	5.	0.83	0.17
8	17.	13.	4.	0.76	0.24
9	7.	5.	2.	0.71	0.29
10	6.	4.	2.	0.67	0.33
11	0.	0.	0.	0.	0.
12	2.	2.	0.	1.00	0.
13	0.	0.	0.	0.	0.
14	0.	0.	0.	0.	0.
15	0.	0.	0.	0.	0.
16	0.	0.	0.	0.	0.
17	0.	0.	0.	0.	0.
18	0.	0.	0.	0.	0.
19	0.	0.	0.	0.	0.
20	0.	0.	0.	0.	0.
TOTAL	255.	205.	50.	0.80	0.20

	MEAN	S.D.
CORRECT DOCUMENTS	4.43	2.52
INCORRECT DOCUMENTS	4.72	2.42

TABLE 3 EFFECTIVENESS VS WORDS IN INTERSECTION OF DOCUMENTS AND DISCRIMINATING SET
70 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 7 48 DISCRIMINATING WORDS

TEST DOCUMENTS			CATEGORY P3		
NUMBER OF WORDS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0	1.	0.	1.	0.	1.00
1	4.	3.	1.	0.75	0.25
2	15.	12.	3.	0.80	0.20
3	26.	22.	4.	0.85	0.15
4	40.	38.	2.	0.95	0.05
5	31.	29.	2.	0.94	0.06
6	20.	18.	2.	0.90	0.10
7	23.	22.	1.	0.96	0.04
8	13.	11.	2.	0.85	0.15
9	11.	9.	2.	0.82	0.18
10	9.	8.	1.	0.89	0.11
11	3.	3.	0.	1.00	0.
12	0.	0.	0.	0.	0.
13	0.	0.	0.	0.	0.
14	0.	0.	0.	0.	0.
15	0.	0.	0.	0.	0.
16	0.	0.	0.	0.	0.
17	0.	0.	0.	0.	0.
18	0.	0.	0.	0.	0.
19	0.	0.	0.	0.	0.
20	0.	0.	0.	0.	0.
TOTAL	196.	175.	21.	0.89	0.11

	MEAN	S.D.
CORRECT DOCUMENTS	5.34	2.29
INCORRECT DOCUMENTS	4.76	2.81

TABLE 3 EFFECTIVENESS VS WORDS IN INTERSECTION OF DOCUMENTS AND DISCRIMINATING SET
70 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 7 48 DISCRIMINATING WORDS

TEST DOCUMENTS			CATEGORY P4		
NUMBER OF WORDS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0	0.	0.	0.	0.	0.
1	6.	3.	3.	0.50	0.50
2	10.	7.	3.	0.70	0.30
3	12.	10.	2.	0.83	0.17
4	11.	10.	1.	0.91	0.09
5	9.	8.	1.	0.89	0.11
6	10.	7.	3.	0.70	0.30
7	8.	8.	0.	1.00	0.
8	4.	4.	0.	1.00	0.
9	1.	1.	0.	1.00	0.
10	0.	0.	0.	0.	0.
11	4.	4.	0.	1.00	0.
12	1.	1.	0.	1.00	0.
13	0.	0.	0.	0.	0.
14	0.	0.	0.	0.	0.
15	0.	0.	0.	0.	0.
16	0.	0.	0.	0.	0.
17	0.	0.	0.	0.	0.
18	0.	0.	0.	0.	0.
19	0.	0.	0.	0.	0.
20	0.	0.	0.	0.	0.
TOTAL	76.	63.	13.	0.83	0.17

	MEAN	S.D.
CORRECT DOCUMENTS	5.11	2.66
INCORRECT DOCUMENTS	3.23	1.83

TABLE 3 EFFECTIVENESS VS WORDS IN INTERSECTION OF DOCUMENTS AND DISCRIMINATING SET
70 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 7 48 DISCRIMINATING WORDS

TEST DOCUMENTS			CATEGORY TOTAL		
NUMBER OF WORDS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0	7.	6.	1.	0.86	0.14
1	33.	28.	5.	0.85	0.15
2	57.	40.	17.	0.70	0.30
3	80.	66.	14.	0.82	0.17
4	85.	76.	9.	0.89	0.11
5	73.	64.	9.	0.88	0.12
6	53.	43.	10.	0.81	0.19
7	61.	55.	6.	0.90	0.10
8	34.	28.	6.	0.82	0.18
9	19.	15.	4.	0.79	0.21
10	15.	12.	3.	0.80	0.20
11	7.	7.	0.	1.00	0.
12	3.	3.	0.	1.00	0.
13	0.	0.	0.	0.	0.
14	0.	0.	0.	0.	0.
15	0.	0.	0.	0.	0.
16	0.	0.	0.	0.	0.
17	0.	0.	0.	0.	0.
18	0.	0.	0.	0.	0.
19	0.	0.	0.	0.	0.
20	0.	0.	0.	0.	0.
TOTAL	527.	443.	84.	0.84	0.16

	MEAN	S.D.
CORRECT DOCUMENTS	4.89	2.49
INCORRECT DOCUMENTS	4.50	2.51

TABLE 4 EFFECTIVENESS VS RADIUS
70 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 7 48 DISCRIMINATING WORDS

TEST DOCUMENTS			CATEGORY P2		
RADIUS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0. - 0.49	125.	94.	31.	0.75	0.25
0.5 - 0.99	76.	62.	14.	0.82	0.18
1.0 - 1.99	45.	40.	5.	0.89	0.11
2.0 - 2.99	7.	7.	0.	1.00	0.
3.0 - 3.99	2.	2.	0.	1.00	0.
4.0 - 4.99	0.	0.	0.	0.	0.
5.0 - 5.99	0.	0.	0.	0.	0.
6.0 - 6.99	0.	0.	0.	0.	0.
7.0 - 7.99	0.	0.	0.	0.	0.
8.0 - 8.99	0.	0.	0.	0.	0.
9.0 - 9.99	0.	0.	0.	0.	0.
10. - UP	0.	0.	0.	0.	0.
TOTAL	255.	205.	50.	0.80	0.20

	MEAN	S.D.
CORRECT DOCUMENTS	0.69	0.60
INCORRECT DOCUMENTS	0.49	0.39

TABLE 4 EFFECTIVENESS VS RADIUS
70 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 7 48 DISCRIMINATING WORDS

TEST DOCUMENTS			CATEGORY P3		
RADIUS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0. - 0.49	47.	33.	14.	0.70	0.30
0.5 - 0.99	67.	64.	3.	0.96	0.04
1.0 - 1.99	66.	62.	4.	0.94	0.06
2.0 - 2.99	14.	14.	0.	1.00	0.
3.0 - 3.99	2.	2.	0.	1.00	0.
4.0 - 4.99	0.	0.	0.	0.	0.
5.0 - 5.99	0.	0.	0.	0.	0.
6.0 - 6.99	0.	0.	0.	0.	0.
7.0 - 7.99	0.	0.	0.	0.	0.
8.0 - 8.99	0.	0.	0.	0.	0.
9.0 - 9.99	0.	0.	0.	0.	0.
10. - UP	0.	0.	0.	0.	0.
TOTAL	196.	175.	21.	0.89	0.11

	MEAN	S.D.
CORRECT DOCUMENTS	1.02	0.63
INCORRECT DOCUMENTS	0.49	0.42

TABLE 4 EFFECTIVENESS VS RADIUS
70 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 7 48 DISCRIMINATING WORDS

TEST DOCUMENTS		CATEGORY P4			
RADIUS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0. - 0.49	24.	14.	10.	0.58	0.42
0.5 - 0.99	28.	25.	3.	0.89	0.11
1.0 - 1.99	20.	20.	0.	1.00	0.
2.0 - 2.99	3.	3.	0.	1.00	0.
3.0 - 3.99	1.	1.	0.	1.00	0.
4.0 - 4.99	0.	0.	0.	0.	0.
5.0 - 5.99	0.	0.	0.	0.	0.
6.0 - 6.99	0.	0.	0.	0.	0.
7.0 - 7.99	0.	0.	0.	0.	0.
8.0 - 8.99	0.	0.	0.	0.	0.
9.0 - 9.99	0.	0.	0.	0.	0.
10. - UP	0.	0.	0.	0.	0.
TOTAL	76.	63.	13.	0.83	0.17

	MEAN	S.D.
CORRECT DOCUMENTS	1.00	0.68
INCORRECT DOCUMENTS	0.34	0.23

TABLE 4 EFFECTIVENESS VS RADIUS
70 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 7 48 DISCRIMINATING WORDS

	TEST DOCUMENTS		CATEGORY TOTAL		
RADIUS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0. - 0.49	196.	141.	55.	0.72	0.28
0.5 - 0.99	171.	151.	20.	0.88	0.12
1.0 - 1.99	131.	122.	9.	0.93	0.07
2.0 - 2.99	24.	24.	0.	1.00	0.
3.0 - 3.99	5.	5.	0.	1.00	0.
4.0 - 4.99	0.	0.	0.	0.	0.
5.0 - 5.99	0.	0.	0.	0.	0.
6.0 - 6.99	0.	0.	0.	0.	0.
7.0 - 7.99	0.	0.	0.	0.	0.
8.0 - 8.99	0.	0.	0.	0.	0.
9.0 - 9.99	0.	0.	0.	0.	0.
10. - UP	0.	0.	0.	0.	0.
TOTAL	527.	443.	84.	0.84	0.16

	MEAN	S.D.
CORRECT DOCUMENTS	0.86	0.64
INCORRECT DOCUMENTS	0.47	0.38

```
TABLE 5   DOCUMENT CLASSIFICATION SUMMARY
          70 DOCUMENTS IN EACH CATEGORY   DISCRIMINATING WORD SET    7     48 DISCRIMINATING  WORDS

  TEST
               AUTO   CATEGORY
ACTUAL
CATEGORY      P2       P3       P4       TOTAL

P2           205.00    17.00    33.00   255.00

P3             9.00   175.00    12.00   196.00

P4   .        12.00     1.00    63.00    76.00

TOTAL        226.00   193.00   108.00   527.00

                 PERCENTAGE

P2             0.80     0.07     0.13     1.00

P3             0.05     0.89     0.06     1.00

P4             0.16     0.01     0.83     1.00

                       SWETS MEASURES               RECALL          RELEVANCE        PRECISION
CATEGORY P2        PERTINENT   NOT PERTINENT        RATIO             RATIO            RATIO
 RETRIEVED           0.80           0.04             0.80              0.91             0.89


                       SWETS MEASURES               RECALL          RELEVANCE        PRECISION
CATEGORY P3        PERTINENT   NOT PERTINENT        RATIO             RATIO            RATIO
 RETRIEVED           0.89           0.03             0.39              0.91             0.98


                       SWETS MEASURES               RECALL          RELEVANC         PRECISION
CATEGORY P4        PERTINENT   NOT PERTINENT        RATIO             RATIO            RATIO
 RETRIEVED           0.83           0.07             0.83              0.58             1.42
```

TABLE 1 EFFECTIVENESS VS NUMBER OF SENTENCES
140 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 7 48 DISCRIMINATING WORDS

SAMPLE DOCUMENTS			CATEGORY P2		
NUMBER OF SENTENCES	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
1	2.	2.	0.	1.00	0.
2	13.	11.	2.	0.85	0.15
3	24.	22.	2.	0.92	0.08
4	15.	14.	1.	0.93	0.07
5	30.	29.	1.	0.97	0.03
6	24.	23.	1.	0.96	0.04
7	10.	10.	0.	1.00	0.
8	11.	10.	1.	0.91	0.09
9	5.	4.	1.	0.80	0.20
10	3.	3.	0.	1.00	0.
11	1.	1.	0.	1.00	0.
12	1.	1.	0.	1.00	0.
13	0.	0.	0.	0.	0.
14	0.	0.	0.	0.	0.
15	0.	0.	0.	0.	0.
TOTAL	139.	130.	9.	0.94	0.06

	MEAN	S.D.
CORRECT DOCUMENTS	5.14	2.16
INCORRECT DOCUMENTS	4.67	2.40

TABLE 1 EFFECTIVENESS VS NUMBER OF SENTENCES
140 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 7 48 DISCRIMINATING WORDS

SAMPLE DOCUMENTS		CATEGORY P3			
NUMBER OF SENTENCES	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
1	1.	1.	0.	1.00	0.
2	11.	9.	2.	0.82	0.18
3	16.	15.	1.	0.94	0.06
4	31.	30.	1.	0.97	0.03
5	19.	19.	0.	1.00	0.
6	22.	20.	2.	0.91	0.09
7	15.	14.	1.	0.93	0.07
8	9.	9.	0.	1.00	0.
9	9.	8.	1.	0.89	0.11
10	3.	3.	0.	1.00	0.
11	2.	2.	0.	1.00	0.
12	1.	1.	0.	1.00	0.
13	0.	0.	0.	0.	0.
14	1.	1.	0.	1.00	0.
15	0.	0.	0.	0.	0.
TOTAL	140.	132.	8.	0.94	0.06

	MEAN	S.D.
CORRECT DOCUMENTS	5.45	2.34
INCORRECT DOCUMENTS	4.88	2.37

TABLE 1 EFFECTIVENESS VS NUMBER OF SENTENCES
140 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 2 49 DISCRIMINATING WORDS

CATEGORY P4

SAMPLE DOCUMENTS NUMBER OF SENTENCES	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
1	2.	2.	0.	1.00	0.
2	13.	9.	4.	0.69	0.31
3	29.	23.	6.	0.79	0.21
4	22.	19.	3.	0.86	0.14
5	21.	15.	6.	0.71	0.29
6	16.	16.	0.	1.00	0.
7	17.	15.	2.	0.86	0.12
8	11.	9.	2.	0.82	0.18
9	2.	1.	1.	0.50	0.50
10	4.	4.	0.	1.00	0.
11	2.	2.	0.	1.00	0.
12	0.	0.	0.	0.	0.
13	0.	0.	0.	0.	0.
14	1.	1.	0.	1.00	0.
15	0.	0.	0.	0.	0.
TOTAL	140.	116.	24.	0.83	0.17

	MEAN	S.D.
CORRECT DOCUMENTS	5.16	2.38
INCORRECT DOCUMENTS	4.46	2.02

TABLE 1 EFFECTIVENESS VS NUMBER OF SENTENCES
140 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 7 48 DISCRIMINATING WORDS

SAMPLE DOCUMENTS			CATEGORY TOTAL		
NUMBER OF SENTENCES	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
1	5.	5.	0.	1.00	0.
2	37.	29.	8.	0.78	0.22
3	69.	60.	9.	0.87	0.13
4	68.	63.	5.	0.93	0.07
5	70.	63.	7.	0.90	0.10
6	62.	59.	3.	0.95	0.05
7	42.	39.	3.	0.93	0.07
8	31.	28.	3.	0.90	0.10
9	16.	13.	3.	0.81	0.19
10	10.	10.	0.	1.00	0.
11	5.	5.	0.	1.00	0.
12	2.	2.	0.	1.00	0.
13	0.	0.	0.	0.	0.
14	2.	2.	0.	1.00	0.
15	0.	0.	0.	0.	0.
TOTAL	419.	378.	41.	0.90	0.10

	MEAN	S.D.
CORRECT DOCUMENTS	5.25	2.29
INCORRECT DOCUMENTS	4.59	2.19

TABLE 2 EFFECTIVENESS VS DOCUMENT LENGTH
140 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 7 48 DISCRIMINATING WORDS

SAMPLF DOCUMENTS			CATEGORY P2		
NUMBER OF TOKENS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0 - 9	0.	0.	0.	0.	0.
10 - 19	0.	0.	0.	0.	0.
20 - 29	2.	2.	0.	1.00	0.
30 - 39	8.	7.	1.	0.88	0.13
40 - 49	9.	8.	1.	0.89	0.11
50 59	7.	5.	2.	0.71	0.29
60 - 69	11.	10.	1.	0.91	0.09
70 - 79	8.	8.	0.	1.00	0.
80 - 89	12.	12.	0.	1.00	0.
90 - 99	12.	12.	0.	1.00	0.
100 - 109	10.	9.	1.	0.90	0.10
110 - 119	9.	9.	0.	1.00	0.
120 - 129	14.	14.	0.	1.00	0.
130 - 139	10.	10.	0.	1.00	0.
140 - 149	4.	4.	0.	1.00	0.
150 - 159	5.	5.	0.	1.00	0.
160 - 169	6.	5.	1.	0.83	0.17
170 - 179	2.	2.	0.	1.00	0.
180 - 189	5.	4.	1.	0.80	0.20
190 - 199	4.	3.	1.	0.75	0.25
200 - 209	0.	0.	0.	0.	0.
210 - 219	0.	0.	0.	0.	0.
220 - 229	0.	0.	0.	0.	0.
230 - 239	1.	1.	0.	1.00	0.
240 - 249	0.	0.	0.	0.	0.
250 - UP	0.	0.	0.	0.	0.
TOTAL	139.	130.	9.	0.94	0.06

	MEAN	S.D.
CORRECT DOCUMENTS	103.58	43.48
INCORRECT DOCUMENTS	101.44	61.24

TABLE /   EFFECTIVENESS VS DOCUMENT LENGTH
140 DOCUMENTS IN EACH CATEGORY   DISCRIMINATING WORD SET   7   48 DISCRIMINATING WORDS

SAMPLE DOCUMENTS NUMBER OF TOKENS		NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	CATEGORY P3 NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0 -	9	0.	0.	0.	0.	0.
10 -	19	0.	0.	0.	0.	0.
20 -	29	0.	0.	0.	0.	0.
30 -	39	6.	5.	1.	0.83	0.17
40 -	49	5.	5.	0.	1.00	0.
50 -	59	8.	7.	1.	0.88	0.13
60 -	69	11.	11.	0.	1.00	0.
70 -	79	14.	13.	1.	0.93	0.07
80 -	89	11.	10.	1.	0.91	0.09
90 -	99	10.	9.	1.	0.90	0.10
100 -	109	10.	10.	0.	1.00	0.
110 -	119	11.	9.	2.	0.82	0.18
120 -	129	10.	10.	0.	1.00	0.
130 -	139	9.	8.	1.	0.89	0.11
140 -	149	8.	8.	0.	1.00	0.
150 -	159	3.	3.	0.	1.00	0.
160 -	169	4.	4.	0.	1.00	0.
170 -	179	6.	6.	0.	1.00	0.
180 -	189	5.	5.	0.	1.00	0.
190 -	199	4.	4.	0.	1.00	0.
200 -	209	2.	2.	0.	1.00	0.
210 -	219	0.	0.	0.	0.	0.
220 -	229	0.	0.	0.	0.	0.
230 -	239	3.	3.	0.	1.00	0.
240 -	249	0.	0.	0.	0.	0.
250 -	UP	0.	0.	0.	0.	0.
TOTAL		140.	132.	8.	0.94	0.06

	MEAN	S.D.
CORRECT DOCUMENTS	110.95	47.56
INCORRECT DOCUMENTS	89.25	34.73

TABLE 2 EFFECTIVENESS VS DOCUMENT LENGTH
140 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 7 48 DISCRIMINATING WORDS

SAMPLE DOCUMENTS		CATEGORY P4			
NUMBER OF TOKENS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0 - 9	0.	0.	0.	0.	0.
10 - 19	0.	0.	0.	0.	0.
20 - 29	1.	0.	1.	0.	1.00
30 - 39	6.	4.	2.	0.67	0.33
40 - 49	6.	5.	1.	0.83	0.17
50 - 59	14.	12.	2.	0.86	0.14
60 - 69	17.	12.	5.	0.71	0.29
70 - 79	15.	14.	1.	0.93	0.07
80 - 89	13.	12.	1.	0.92	0.08
90 - 99	11.	9.	2.	0.82	0.18
100 - 109	10.	7.	3.	0.70	0.30
110 - 119	7.	7.	0.	1.00	0.
120 - 129	9.	8.	1.	0.89	0.11
130 - 139	7.	6.	1.	0.86	0.14
140 - 149	10.	8.	2.	0.80	0.20
150 - 159	4.	3.	1.	0.75	0.25
160 - 169	5.	5.	0.	1.00	0.
170 - 179	1.	1.	0.	1.00	0.
180 - 189	1.	1.	0.	1.00	0.
190 - 199	0.	0.	0.	0.	0.
200 - 209	1.	0.	1.	0.	1.00
210 - 219	1.	1.	0.	1.00	0.
220 - 229	0.	0.	0.	0.	0.
230 - 239	1.	1.	0.	1.00	0.
240 - 249	0.	0.	0.	0.	0.
250 - UP	0.	0.	0.	0.	0.
TOTAL	140.	116.	24.	0.83	0.17

	MEAN	S.D.
CORRECT DOCUMENTS	97.14	40.54
INCORRECT DOCUMENTS	88.54	43.65

TABLE 2 EFFECTIVENESS VS DOCUMENT LENGTH

140 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 7 48 DISCRIMINATING WORDS

SAMPLE DOCUMENTS			CATEGORY TOTAL		
NUMBER OF TOKENS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0 - 9	0.	0.	0.	0.	0.
10 - 19	0.	0.	0.	0.	0.
20 - 29	3.	2.	1.	0.67	0.33
30 - 39	20.	16.	4.	0.80	0.20
40 - 49	20.	18.	2.	0.90	0.10
50 - 59	29.	24.	5.	0.83	0.17
60 - 69	39.	33.	6.	0.85	0.15
70 - 79	37.	35.	2.	0.95	0.05
80 - 89	36.	34.	2.	0.94	0.06
90 - 99	33.	30.	3.	0.91	0.09
100 - 109	30.	26.	4.	0.87	0.13
110 - 119	27.	25.	2.	0.93	0.07
120 - 129	33.	32.	1.	0.97	0.03
130 - 139	26.	24.	2.	0.92	0.08
140 - 149	22.	20.	2.	0.91	0.09
150 - 159	12.	11.	1.	0.92	0.08
160 - 169	15.	14.	1.	0.93	0.07
170 - 179	9.	9.	0.	1.00	0.
180 - 189	11.	10.	1.	0.91	0.09
190 - 199	8.	7.	1.	0.88	0.13
200 - 209	3.	2.	1.	0.67	0.33
210 - 219	1.	1.	0.	1.00	0.
220 - 229	0.	0.	0.	0.	0.
230 - 239	5.	5.	0.	1.00	0.
240 - 249	0.	0.	0.	0.	0.
250 - UP	0.	0.	0.	0.	0.
TOTAL	419.	378.	41.	0.90	0.10

	MEAN	S.D.
CORRECT DOCUMENTS	104.18	44.45
INCORRECT DOCUMENTS	91.51	46.92

TABLE 3 EFFECTIVENESS VS WORDS IN INTERSECTION OF DOCUMENTS AND DISCRIMINATING SET
140 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 7 48 DISCRIMINATING WORDS

SAMPLE DOCUMENTS			CATEGORY P2		
NUMBER OF WORDS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0	1.	1.	0.	1.00	0.
1	13.	13.	0.	1.00	0.
2	13.	11.	2.	0.85	0.15
3	28.	25.	3.	0.89	0.11
4	17.	16.	1.	0.94	0.06
5	22.	21.	1.	0.95	0.05
6	16.	16.	0.	1.00	0.
7	12.	11.	1.	0.92	0.08
8	11.	10.	1.	0.91	0.09
9	4.	4.	0.	1.00	0.
10	2.	2.	0.	1.00	0.
11	0.	0.	0.	0.	0.
12	0.	0.	0.	0.	0.
13	0.	0.	0.	0.	0.
14	0.	0.	0.	0.	0.
15	0.	0.	0.	0.	0.
16	0.	0.	0.	0.	0.
17	0.	0.	0.	0.	0.
18	0.	0.	0.	0.	0.
19	0.	0.	0.	0.	0.
20	0.	0.	0.	0.	0.
TOTAL	139.	130.	9.	0.94	0.06

	MEAN	S.D.
CORRECT DOCUMENTS	4.52	2.29
INCORRECT DOCUMENTS	4.11	2.02

TABLE 3 EFFECTIVENESS VS WORDS IN INTERSECTION OF DOCUMENTS AND DISCRIMINATING SET
140 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 7 48 DISCRIMINATING WORDS

SAMPLE DOCUMENTS			CATEGORY P3		
NUMBER OF WORDS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENIS	PERCENTAGE OF INCORRECT DOCUMENTS
0	0.	0.	0.	0.	0.
1	2.	1.	1.	0.50	0.50
2	11.	9.	2.	0.82	0.18
3	14.	12.	2.	0.86	0.14
4	22.	22.	0.	1.00	0.
5	20.	18.	2.	0.90	0.10
6	20.	20.	0.	1.00	0.
7	17.	17.	0.	1.00	0.
8	14.	14.	0.	1.00	0.
9	8.	7.	1.	0.88	0.13
10	9.	9.	0.	1.00	0.
11	2.	2.	0.	1.00	0.
12	0.	0.	0.	0.	0.
13	0.	0.	0.	0.	0.
14	1.	1.	0.	1.00	0.
15	0.	0.	0.	0.	0.
16	0.	0.	0.	0.	0.
17	0.	0.	0.	0.	0.
18	0.	0.	0.	0.	0.
19	0.	0.	0.	0.	0.
20	0.	0.	0.	0.	0.
TOTAL	140.	132.	8.	0.94	0.06

	MEAN	S.D.
CORRECT DOCUMENTS	5.86	2.42
INCORRECT DOCUMENTS	3.75	2.38

TABLE 3 EFFECTIVENESS VS WORDS IN INTERSECTION OF DOCUMENTS AND DISCRIMINATING SET
140 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 7 48 DISCRIMINATING WORDS

SAMPLE DOCUMENTS			CATEGORY P4		
NUMBER OF WORDS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0	1.	0.	1.	0.	1.00
1	8.	2.	6.	0.25	0.75
2	17.	11.	6.	0.65	0.35
3	22.	21.	1.	0.95	0.05
4	18.	17.	1.	0.94	0.06
5	21.	19.	2.	0.90	0.10
6	15.	13.	2.	0.87	0.13
7	16.	12.	4.	0.75	0.25
8	8.	8.	0.	1.00	0.
9	7.	7.	0.	1.00	0.
10	2.	1.	1.	0.50	0.50
11	4.	4.	0.	1.00	0.
12	1.	1.	0.	1.00	0.
13	0.	0.	0.	0.	0.
14	0.	0.	0.	0.	0.
15	0.	0.	0.	0.	0.
16	0.	0.	0.	0.	0.
17	0.	0.	0.	0.	0.
18	0.	0.	0.	0.	0.
19	0.	0.	0.	0.	0.
20	0.	0.	0.	0.	0.
TOTAL	140.	116.	24.	0.83	0.17

	MEAN	S.D.
CORRECT DOCUMENTS	5.22	2.44
INCORRECT DOCUMENTS	3.54	2.68

TABLE 3 EFFECTIVENESS VS WORDS IN INTERSECTION OF DOCUMENTS AND DISCRIMINATING SET
140 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 7 48 DISCRIMINATING WORDS

SAMPLE DOCUMENTS			CATEGORY TOTAL		
NUMBER OF WORDS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0	2.	1.	1.	0.50	0.50
1	23.	16.	7.	0.70	0.30
2	41.	31.	10.	0.76	0.24
3	64.	58.	6.	0.91	0.09
4	57.	55.	2.	0.96	0.04
5	63.	58.	5.	0.92	0.08
6	51.	49.	2.	0.96	0.04
7	45.	40.	5.	0.89	0.11
8	33.	32.	1.	0.97	0.03
9	19.	18.	1.	0.95	0.05
10	13.	12.	1.	0.92	0.08
11	6.	6.	0.	1.00	0.
12	1.	1.	0.	1.00	0.
13	0.	0.	0.	0.	0.
14	1.	1.	0.	1.00	0.
15	0.	0.	0.	0.	0.
16	0.	0.	0.	0.	0.
17	0.	0.	0.	0.	0.
18	0.	0.	0.	0.	0.
19	0.	0.	0.	0.	0.
20	0.	0.	0.	0.	0.
TOTAL	419.	378.	41.	0.90	0.10

	MEAN	S.D.
CORRECT DOCUMENTS	5.20	2.45
INCORRECT DOCUMENTS	3.71	2.50

TABLE 4 EFFECTIVENESS VS RADIUS
140 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 7 48 DISCRIMINATING WORDS

SAMPLE DOCUMENTS		CATEGORY P2			
RADIUS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0. - 0.49	55.	50.	5.	0.91	0.09
0.5 - 0.99	51.	50.	1.	0.98	0.02
1.0 - 1.99	29.	26.	3.	0.90	0.10
2.0 - 2.99	4.	4.	0.	1.00	0.
3.0 - 3.99	0.	0.	0.	0.	0.
4.0 - 4.99	0.	0.	0.	0.	0.
5.0 - 5.99	0.	0.	0.	0.	0.
6.0 - 6.99	0.	0.	0.	0.	0.
7.0 - 7.99	0.	0.	0.	0.	0.
8.0 - 8.99	0.	0.	0.	0.	0.
9.0 - 9.99	0.	0.	0.	0.	0.
10. - UP	0.	0.	0.	0.	0.
TOTAL	139.	130.	9.	0.94	0.06

	MEAN	S.D.
CORRECT DOCUMENTS	0.75	0.55
INCORRECT DOCUMENTS	0.68	0.50

TABLE 4 EFFECTIVENESS VS RADIUS

140 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 7 48 DISCRIMINATING WORDS

SAMPLE DOCUMENTS			CATEGORY P3		
RADIUS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0. - 0.49	16.	9.	7.	0.56	0.44
0.5 - 0.99	30.	29.	1.	0.97	0.03
1.0 - 1.99	72.	72.	0.	1.00	0.
2.0 - 2.99	20.	20.	0.	1.00	0.
3.0 - 3.99	2.	2.	0.	1.00	0.
4.0 - 4.99	0.	0.	0.	0.	0.
5.0 - 5.99	0.	0.	0.	0.	0.
6.0 - 6.99	0.	0.	0.	0.	0.
7.0 - 7.99	0.	0.	0.	0.	0.
8.0 - 8.99	0.	0.	0.	0.	0.
9.0 - 9.99	0.	0.	0.	0.	0.
10. - UP	0.	0.	0.	0.	0.
TOTAL	140.	132.	8.	0.94	0.06

	MEAN	S.D.
CORRECT DOCUMENTS	1.35	0.63
INCORRECT DOCUMENTS	0.36	0.25

TABLE 4 EFFECTIVENESS VS RADIUS
140 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 7 48 DISCRIMINATING WORDS

SAMPLE DOCUMENTS			CATEGORY P4		
RADIUS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0. - 0.49	35.	18.	17.	0.51	0.49
0.5 - 0.99	52.	45.	7.	0.87	0.13
1.0 - 1.99	45.	45.	0.	1.00	0.
2.0 - 2.99	8.	8.	0.	1.00	0.
3.0 - 3.99	0.	0.	0.	0.	0.
4.0 - 4.99	0.	0.	0.	0.	0.
5.0 - 5.99	0.	0.	0.	0.	0.
6.0 - 6.99	0.	0.	0.	0.	0.
7.0 - 7.99	0.	0.	0.	0.	0.
8.0 - 8.99	0.	0.	0.	0.	0.
9.0 - 9.99	0.	0.	0.	0.	0.
10. - UP	0.	0.	0.	0.	0.
TOTAL	140.	116.	24.	0.83	0.17

	MEAN	S.D.
CORRECT DOCUMENTS	1.06	0.58
INCORRECT DOCUMENTS	0.35	0.29

TABLE 4 EFFECTIVENESS VS RADIUS
140 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 7 48 DISCRIMINATING WORDS

SAMPLE DOCUMENTS			CATEGORY TOTAL		
RADIUS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0. - 0.49	106.	77.	29.	0.73	0.27
0.5 - 0.99	133.	124.	9.	0.93	0.07
1.0 - 1.99	146.	143.	3.	0.98	0.02
2.0 - 2.99	32.	32.	0.	1.00	0.
3.0 - 3.99	2.	2.	0.	1.00	0.
4.0 - 4.99	0.	0.	0.	0.	0.
5.0 - 5.99	0.	0.	0.	0.	0.
6.0 - 6.99	0.	0.	0.	0.	0.
7.0 - 7.99	0.	0.	0.	0.	0.
8.0 - 8.99	0.	0.	0.	0.	0.
9.0 - 9.99	0.	0.	0.	0.	0.
10. - UP	0.	0.	0.	0.	0.
TOTAL	419.	378.	41.	0.90	0.10

	MEAN	S.D.
CORRECT DOCUMENTS	1.05	0.64
INCORRECT DOCUMENTS	0.43	0.37

```
TABLE 5   DOCUMENT CLASSIFICATION SUMMARY
          140 DOCUMENTS IN EACH CATEGORY   DISCRIMINATING WORD SET    7     48 DISCRIMINATING  WORDS

SAMPLE
                    AUTO   CATEGORY
ACTUAL
CATEGORY         P2        P3        P4        TOTAL

P2               130.00      5.00      4.00   139.00

P3                 6.00    132.00      2.00   140.00

P4                23.00      1.00    116.00   140.00

TOTAL            159.00    138.00    122.00   419.00

                      PERCENTAGE

P2                 0.94      0.04      0.03     1.00

P3                 0.04      0.94      0.01     1.00

P4                 0.16      0.01      0.83     1.00

                          SWETS MEASURES              RECALL        RELEVANCE       PRECISION
CATEGORY P2          PERTINENT    NOT PERTINENT       RATIO           RATIO           RATIO
 RETRIEVED             0.94            0.10            0.94            0.82            1.14


                          SWETS MEASURES              RECALL        RELEVANCE       PRECISION
CATEGORY P3          PERTINENT    NOT PERTINENT       RATIO           RATIO           RATIO
 RETRIEVED             0.94            0.02            0.94            0.96            0.99


                          SWETS MEASURES              RECALL        RELEVANCE       PRECISION
CATEGORY P4          PERTINENT    NOT PERTINENT       RATIO           RATIO           RATIO
 RETRIEVED             0.83            0.02            0.83            0.95            0.87
```

TABLE 1 EFFECTIVENESS VS NUMBER OF SENTENCES
140 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 7 48 DISCRIMINATING WORDS

TEST DOCUMENTS			CATEGORY P2		
NUMBER OF SENTENCES	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
1	0.	0.	0.	0.	0.
2	8.	7.	1.	0.88	0.13
3	29.	26.	3.	0.90	0.10
4	45.	39.	6.	0.37	0.13
5	29.	26.	3.	0.90	0.10
6	26.	20.	6.	0.77	0.23
7	18.	16.	2.	0.89	0.11
8	13.	11.	2.	0.85	0.15
9	9.	8.	1.	0.89	0.11
10	5.	5.	0.	1.00	0.
11	3.	2.	1.	0.67	0.33
12	1.	1.	0.	1.00	0.
13	0.	0.	0.	0.	0.
14	0.	0.	0.	0.	0.
15	0.	0.	0.	0.	0.
TOTAL	186.	161.	25.	0.87	0.13

	MEAN	S.D.
CORRECT DOCUMENTS	5.30	2.16
INCORRECT DOCUMENTS	5.44	2.06

TABLE 1 EFFECTIVENESS VS NUMBER OF SENTENCES
140 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 7 48 DISCRIMINATING WORDS

TEST DOCUMENTS			CATEGORY P3		
NUMBER OF SENTENCES	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
1	3.	1.	2.	0.33	0.67
2	7.	6.	1.	0.86	0.14
3	15.	14.	1.	0.93	0.07
4	24.	23.	1.	0.96	0.04
5	24.	23.	1.	0.96	0.04
6	17.	15.	2.	0.88	0.12
7	15.	14.	1.	0.93	0.07
8	5.	4.	1.	0 80	0.20
9	9.	9.	0.	1.00	0.
10	5.	5.	0.	1.00	0.
11	1.	1.	0.	1.00	0.
12	1.	1.	0.	1.00	0.
13	0.	0.	0.	0.	0.
14	0.	0.	0.	0.	0.
15	0.	0.	0.	0.	0.
TOTAL	126.	116.	10.	0.92	0.08

	MEAN	S.D.
CORRECT DOCUMENTS	5.48	2.24
INCORRECT DOCUMENTS	4.30	2.37

TABLE 1 EFFECTIVENESS VS NUMBER OF SENTENCES
140 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 7 48 DISCRIMINATING WORDS

TEST DOCUMENTS			CATEGORY P4		
NUMBER OF SENTENCES	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
1	2.	2.	0.	1.00	0.
2	15.	12.	3.	0.80	0.20
3	28.	22.	6.	0.79	0.21
4	22.	20.	2.	0.91	0.09
5	22.	16.	6.	0.73	0.27
6	16.	16.	0.	1.00	0.
7	16.	15.	1.	0.94	0.06
8	11.	9.	2.	0.82	0.18
9	2.	1.	1.	0.50	0.50
10	4.	4.	0.	1.00	0.
11	2.	2.	0.	1.00	0.
12	0.	0.	0.	0.	0.
13	0.	0.	0.	0.	0.
14	1.	1.	0.	1.00	0.
15	0.	0.	0.	0.	0.
TOTAL	141.	120.	21.	0.85	0.15

	MEAN	S.D.
CORRECT DOCUMENTS	5.08	2.37
INCORRECT DOCUMENTS	4.48	2.01

TABLE 1 EFFECTIVENESS VS NUMBER OF SENTENCES
140 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 7 48 DISCRIMINATING WORDS

TEST DOCUMENTS			CATEGORY TOTAL		
NUMBER OF SENTENCES	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORFECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
1	5.	3.	2.	0.60	0.40
2	30.	25.	5.	0.83	0.17
3	72.	62.	10.	0.86	0.14
4	91.	82.	9.	0.90	0.10
5	75.	65.	10.	0.87	0.13
6	59.	51.	8.	0.86	0.14
7	49.	45.	4.	0.92	0.08
8	29.	24.	5.	0.83	0.17
9	20.	18.	2.	0.90	0.10
10	14.	14.	0.	1.00	0.
11	6.	5.	1.	0.83	0.17
12	2.	2.	0.	1.00	0.
13	0.	0.	0.	0.	0.
14	1.	1.	0.	1.00	0.
15	0.	0.	0.	0.	0.
TOTAL	453.	397.	56.	0.88	0.12

	MEAN	S.D.
CORRECT DOCUMENTS	5.29	2.25
INCORRECT DOCUMENTS	4.88	2.16

TABLE 2 EFFECTIVENESS VS DOCUMENT LENGTH

140 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 7 48 DISCRIMINATING WORDS

TEST DOCUMENTS		CATEGORY P2			
NUMBER OF TOKENS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0 - 9	0.	0.	0.	0.	0.
10 - 19	0.	0.	0.	0.	0.
20 - 29	0.	0.	0.	0.	0.
30 - 39	5.	3.	2.	0.60	0.40
40 - 49	8.	8.	0.	1.00	0.
50 - 59	19.	17.	2.	0.89	0.11
60 - 69	22.	20.	2.	0.91	0.09
70 - 79	20.	17.	3.	0.85	0.15
80 - 89	17.	15.	2.	0.88	0.12
90 - 99	11.	9.	2.	0.82	0.18
100 - 109	10.	9.	1.	0.90	0.10
110 - 119	9.	9.	0.	1.00	0.
120 - 129	15.	12.	3.	0.80	0.20
130 - 139	8.	7.	1.	0.88	0.13
140 - 149	9.	8.	1.	0.89	0.11
150 - 159	9.	8.	1.	0.89	0.11
160 - 169	5.	4.	1.	0.80	0.20
170 - 179	11.	9.	2.	0.82	0.18
180 - 189	2.	1.	1.	0.50	0.50
190 - 199	0.	0.	0.	0.	0.
200 - 209	2.	2.	0.	1.00	0.
210 - 219	0.	0.	0.	0.	0.
220 - 229	1.	0.	1.	0.	1.00
230 - 239	2.	2.	0.	1.00	0.
240 - 249	0.	0.	0.	0.	0.
250 - UP	1.	1.	0.	1.00	0.
TOTAL	186.	161.	25.	0.87	0.13

	MEAN	S.D.
CORRECT DOCUMENTS	101.59	45.32
INCORRECT DOCUMENTS	108.68	49,46

TABLE 2 EFFECTIVENESS VS DOCUMENT LENGTH
140 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 7 48 DISCRIMINATING WORDS

TEST DOCUMENTS			CATEGORY P3		
NUMBER OF TOKENS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCFNTAGE OF INCORRECT DOCUMENTS
0 - 9	0.	0.	0.	0.	0.
10 - 19	0.	0.	0.	0.	0.
20 - 29	2.	1.	1.	0.50	0.50
30 - 39	3.	3.	0.	1.00	0.
40 - 49	9.	7.	2.	0.78	0.22
50 - 59	8.	6.	2.	0.75	0.25
60 - 69	11.	11.	0.	1.00	0.
70 - 79	9.	8.	1.	0.89	0.11
80 - 89	7.	7.	0.	1.00	0.
90 - 99	13.	12.	1.	0.92	0.08
100 - 109	11.	11.	0.	1.00	0.
110 - 119	10.	10.	0.	1.00	0.
120 - 129	12.	11.	1.	0.92	0.08
130 - 139	4.	3.	1.	0.75	0.25
140 - 149	1.	1.	0.	1.00	0.
150 - 159	4.	4.	0.	1.00	0.
160 - 169	6.	5.	1.	0.83	0.17
170 - 179	6.	6.	0.	1.00	0.
180 - 189	4.	4.	0.	1.00	0.
190 - 199	2.	2.	0.	1.00	0.
200 - 209	1.	1.	0.	1.00	0.
210 - 219	2.	2.	0.	1.00	0.
220 - 229	0.	0.	0.	0.	0.
230 - 239	0.	0.	0.	0.	0.
240 - 249	1.	1.	0.	1.00	0.
250 - UP	0.	0.	0.	0.	0.
TOTAL	126.	116.	10.	0.92	0.08

	MEAN	S.D.
CORRECT DOCUMENTS	108.16	45,88
INCORRECT DOCUMENTS	82.10	44.34

TABLE 2 EFFECTIVENESS VS DOCUMENT LENGTH
140 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 7 48 DISCRIMINATING WORDS

TEST DOCUMENTS		CATEGORY P4			
NUMBER OF TOKENS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0 - 9	0.	0.	0.	0.	0.
10 - 19	0.	0.	0.	0.	0.
20 - 29	0.	0.	0.	0.	0.
30 - 39	8.	6.	2.	0.75	0.25
40 - 49	6.	5.	1.	0.83	0.17
50 - 59	14.	11.	3.	0.79	0.21
60 - 69	19.	15.	4.	0.79	0.21
70 - 79	15.	14.	1.	0.93	0.07
80 - 89	12.	12.	0.	1.00	0.
90 - 99	11.	9.	2.	0.82	0.18
100 - 109	10.	7.	3.	0.70	0.30
110 - 119	8.	8.	0.	1.00	0.
120 - 129	8.	7.	1.	0.88	0.13
130 - 139	7.	6.	1.	0.86	0.14
140 - 149	10.	8.	2.	0.80	0.20
150 - 159	3.	3.	0.	1.00	0.
160 - 169	5.	5.	0.	1.00	0.
170 - 179	1.	1.	0.	1.00	0.
180 - 189	1.	1.	0.	1.00	0.
190 - 199	0.	0.	0.	0.	0.
200 - 209	1.	0.	1.	0.	1.00
210 - 219	1.	1.	0.	1.00	0.
220 - 229	0.	0.	0.	0.	0.
230 - 239	1.	1.	0.	1.00	0.
240 - 249	0.	0.	0.	0.	0.
250 - UP	0.	0.	0.	0.	0.
TOTAL	141.	120.	21.	0.85	0.15

	MEAN	S.D.
CORRECT DOCUMENTS	95.58	40.78
INCORRECT DOCUMENTS	88.24	42.49

TABLE 2 EFFECTIVENESS VS DOCUMENT LENGTH
140 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 7 48 DISCRIMINATING WORDS

TEST DOCUMENTS			CATEGORY TOTAL		
NUMBER OF TOKENS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0 - 9	0.	0.	0.	0.	0.
10 - 19	0.	0.	0.	0.	0.
20 - 29	2.	1.	1.	0.50	0.50
30 - 39	16.	12.	4.	0.75	0.25
40 - 49	23.	20.	3.	0.87	0.13
50 - 59	41.	34.	7.	0.83	0.17
60 - 69	52.	46.	6.	0.88	0.12
70 - 79	44.	39.	5.	0.89	0.11
80 - 89	36.	34.	2.	0.94	0.06
90 - 99	35.	30.	5.	0.86	0.14
100 - 109	31.	27.	4.	0.37	0.13
110 - 119	27.	27.	0.	1.00	0.
120 - 129	35.	30.	5.	0.86	0.14
130 - 139	19.	16.	3.	0.84	0.16
140 - 149	20.	17.	3.	0.85	0.15
150 - 159	16.	15.	1.	0.94	0.06
160 - 169	16.	14.	2.	0.88	0.13
170 - 179	18.	16.	2.	0.89	0.11
180 - 189	7.	6.	1.	0.86	0.14
190 - 199	2.	2.	0.	1.00	0.
200 - 209	4.	3.	1.	0.75	0.25
210 - 219	3.	3.	0.	1.00	0.
220 - 229	1.	0.	1.	0.	1.00
230 - 239	3.	3.	0.	1.00	0.
240 - 249	1.	1.	0.	1.00	0.
250 - UP	1.	1.	0.	1.00	0.
TOTAL	453.	397.	56.	0.88	0.12

	MEAN	S.D.
CORRECT DOCUMENTS	101.70	44.43
INCORRECT DOCUMENTS	96.27	47.42

TABLE 3 EFFECTIVENESS VS WORDS IN INTERSECTION OF DOCUMENTS AND DISCRIMINATING SET
140 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 7 48 DISCRIMINATING WORDS

TEST DOCUMENTS		CATEGORY P2			
NUMBER OF WORDS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0	5.	5.	0.	1.00	0.
1	14.	14.	0.	1.00	0.
2	27.	23.	4.	0.85	0.15
3	26.	24.	2.	0.92	0.08
4	23.	21.	2.	0.91	0.09
5	23.	17.	6.	0.74	0.26
6	17.	13.	4.	0.76	0.24
7	29.	26.	3.	0.90	0.10
8	10.	9.	1.	0.90	0.10
9	5.	4.	1.	0.80	0.20
10	5.	4.	1.	0.80	0.20
11	0.	0.	0.	0.	0.
12	2.	1.	1.	0.50	0.50
13	0.	0.	0.	0.	0.
14	0.	0.	0.	0.	0.
15	0.	0.	0.	0.	0.
16	0.	0.	0.	0.	0.
17	0.	0.	0.	0.	0.
18	0.	0.	0.	0.	0.
19	0.	0.	0.	0.	0.
20	0.	0.	0.	0.	0.
TOTAL	186.	161,	25.	0.87	0.13

	MEAN	S.D.
CORRECT DOCUMENTS	4.48	2.54
INCORRECT DOCUMENTS	5.44	2.50

TABLE 3 EFFECTIVENESS VS WORDS IN INTERSECTION OF DOCUMENTS AND DISCRIMINATING SET
140 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 7 48 DISCRIMINATING WORDS

TEST DOCUMENTS			CATEGORY P3		
NUMBER OF WORDS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0	1.	0.	1.	0.	1.00
1	3.	2.	1.	0.67	0.33
2	10.	7.	3.	0.70	0.30
3	18.	17.	1.	0.94	0.06
4	30.	28.	2.	0.93	0.07
5	20.	18.	2.	0.90	0.10
6	10.	10.	0.	1.00	0.
7	15.	15.	0.	1.00	0.
8	7.	7.	0.	1.00	0.
9	9.	9.	0.	1.00	0.
10	2.	2.	0.	1.00	0.
11	1.	1.	0.	1.00	0.
12	0.	0.	0.	0.	0.
13	0.	0.	0.	0.	0.
14	0.	0.	0.	0.	0.
15	0.	0.	0.	0.	0.
16	0.	0.	0.	0.	0.
17	0.	0.	0.	0.	0.
18	0.	0.	0.	0.	0.
19	0.	0.	0.	0.	0.
20	0.	0.	0.	0.	0.
TOTAL	126.	116.	10.	0.92	0.08

	MEAN	S.D.
CORRECT DOCUMENTS	5.19	2.18
INCORRECT DOCUMENTS	2.80	1.60

TABLE 3 EFFLCTIVENESS VS WORDS IN INTERSECTION OF DOCUMENTS AND DISCRIMINATING SET
140 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 7 48 DISCRIMINATING WORDS

TEST DOCUMENTS			CATEGORY P4		
NUMBER OF WORDS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0	0.	0.	0.	0.	0.
1	7.	3.	4.	0.43	0.57
2	19.	12.	7.	0.63	0.37
3	20.	19.	1.	0.95	0.05
4	17.	16.	1.	0.94	0.06
5	21.	19.	2.	0.90	0.10
6	19.	17.	2.	0.89	0.11
7	15.	13.	2.	0.87	0.13
8	9.	8.	1.	0.89	0.11
9	7.	7.	0.	1.00	0.
10	2.	1.	1.	0.50	0.50
11	4.	4.	0.	1.00	0.
12	1.	1.	0.	1.00	0.
13	0.	0.	0.	0.	0.
14	0.	0.	0.	0.	0.
15	0.	0.	0.	0.	0.
16	0.	0.	0.	0.	0.
17	0.	0.	0.	0.	0.
18	0.	0.	0.	0.	0.
19	0.	0.	0.	0.	0.
20	0.	0.	0.	0.	0.
TOTAL	141.	120.	21.	0.85	0.15

	MEAN	S.D.
CORRECT DOCUMENTS	5.24	2.44
INCORRECT DOCUMENTS	3.76	2.62

TABLE 3   EFFECTIVENESS VS WORDS IN INTERSECTION OF DOCUMENTS AND DISCRIMINATING SET
140 DOCUMENTS IN EACH CATEGORY   DISCRIMINATING WORD SET   7   48 DISCRIMINATING WORDS

TEST DOCUMENTS			CATEGORY TOTAL		
NUMBER OF WORDS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0	6.	5.	1.	0.83	0.17
1	24.	19.	5.	0.79	0.21
2	56.	42.	14.	0.75	0.25
3	64.	60.	4.	0.94	0.06
4	70.	65.	5.	0.93	0.07
5	64.	54.	10.	0.84	0.16
6	46.	40.	6.	0.87	0.13
7	59.	54.	5.	0.92	0.08
8	26.	24.	2.	0.92	0.08
9	21.	20.	1.	0.95	0.05
10	9.	7.	2.	0.78	0.22
11	5.	5.	0.	1.00	0.
12	3.	2.	1.	0.67	0.33
13	0.	0.	0.	0.	0.
14	0.	0.	0.	0.	0.
15	0.	0.	0.	0.	0.
16	0.	0.	0.	0.	0.
17	0.	0.	0.	0.	0.
18	0.	0.	0.	0.	0.
19	0.	0.	0.	0.	0.
20	0.	0.	0.	0.	0.
TOTAL	453.	397.	56.	0.88	0.12

	MEAN	S.D.
CORRECT DOCUMENTS	4.92	2.44
INCORRECT DOCUMENTS	4.34	2.63

TABLE 4 EFFECTIVENESS VS RADIUS
140 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 7 48 DISCRIMINATING WORDS

	TEST DOCUMENTS		CATEGORY P2		
RADIUS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0. - 0.49	84.	74.	10.	0.88	0.12
0.5 - 0.99	60.	50.	10.	0.83	0.17
1.0 - 1.99	35.	32.	3.	0.91	0.09
2.0 - 2.99	4.	3.	1.	0.75	0.25
3.0 - 3.99	2.	1.	1.	0.50	0.50
4.0 - 4.99	1.	1.	0.	1.00	0.
5.0 - 5.99	0.	0.	0.	0.	0.
6.0 - 6.99	0.	0.	0.	0.	0.
7.0 - 7.99	0.	0.	0.	0.	0.
8.0 - 8.99	0.	0.	0.	0.	0.
9.0 - 9.99	0.	0.	0.	0.	0.
10. - UP	0.	0.	0.	0.	0.
TOTAL	186.	161.	25.	0.87	0.13

	MEAN	S.D.
CORRECT DOCUMENTS	0.71	0.62
INCORRECT DOCUMENTS	0.80	0.70

```
TABLE 4   EFFECTIVENESS VS RADIUS
          140 DOCUMENTS IN EACH CATEGORY   DISCRIMINATING WORD SET    7     48 DISCRIMINATING  WORDS

          TEST DOCUMENTS                          CATEGORY P3
RADIUS         NUMBER OF    NUMBER OF               NUMBER OF                PERCENTAGE OF          PERCENTAGE OF
               DOCUMENTS     CORRECT DOCUMENTS       INCORRECT DOCUMENTS     CORRECT DOCUMENTS      INCORRECT DOCUMENTS
0.   - 0.49    13.              6.                      7.                       0.46                   0.54

0.S - 0.99     39.             36.                      3.                       0.92                   0.08

1.0 - 1.99     57.             57.                      0.                       1.00                   0.

2.0 - 2.99     12.             12.                      0.                       1.00                   9.

3.0 - 3.99      4.              4.                      0.                       1.00                   0.

4.0 - 4.99      1.              1.                      0.                       1.00                   0.

5.0 - 5.99      0.              0.                      0.                       0.                     0.

6.0 - 6.99      0.              0.                      0.                       0.                     0.

7.0 - 7.99      0.              0.                      0.                       0.                     0.

8.0 - 8.99      0.              0.                      0.                       0.                     0.

9.0 - 9.99      0.              0.                      0.                       0.                     0.

10. -      UP   0.              0.                      0.                       0.                     0.

TOTAL          126.            116.                    10.                       0.92                   0.08

                          MEAN               S.D.
CORRECT   DOCUMENTS        1.32               0.73
INCORRECT DOCUMENTS        0.36               0.30
```

TABLE 4 EFFECTIVENESS VS RADIUS
140 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 7 48 DISCRIMINATING WORDS

	TEST DOCUMENTS		CATEGORY P4		
RADIUS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DOCUMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0. - 0.49	34.	19.	15.	0.56	0.44
0.5 - 0.99	47.	42.	5.	0.89	0.11
1.0 - 1.99	52.	51.	1.	0.98	0.02
2.0 - 2.99	8.	8.	0.	1.00	0.
3.0 - 3.99	0.	0.	0.	0.	0.
4.0 - 4.99	0.	0.	0.	0.	0.
5.0 - 5.99	0.	0.	0.	0.	0.
6.0 - 6.99	0.	0.	0.	0.	0.
7.0 - 7.99	0.	0.	0.	0.	0.
8.0 - 8.99	0.	0.	0.	0.	0.
9.0 - 9.99	0.	0.	0.	0.	0.
10. - UP	0.	0.	0.	0.	0.
TOTAL	141.	120.	21.	0.85	0.15

	MEAN	S.D.
CORRECT DOCUMENTS	1.07	0.58
INCORRECT DOCUMENTS	0.40	0.39

TABLE 4 EFFECTIVENESS VS RADIUS
140 DOCUMENTS IN EACH CATEGORY DISCRIMINATING WORD SET 7 48 DISCRIMINATING WORDS

	TEST DOCUMENTS		CATEGORY TOTAL		
RADIUS	NUMBER OF DOCUMENTS	NUMBER OF CORRECT DO^UMENTS	NUMBER OF INCORRECT DOCUMENTS	PERCENTAGE OF CORRECT DOCUMENTS	PERCENTAGE OF INCORRECT DOCUMENTS
0. - 0.49	131.	99.	32.	0.76	0.24
0.5 - 0.99	146.	128.	18.	0.88	0.12
1.0 - 1.99	144.	140.	4.	0.97	0.03
2.0 - 2.99	24.	23.	1.	0.96	0.04
3.0 - 3.99	6.	5.	1.	0.83	0.17
4.0 - 4.99	2.	2.	0.	1.00	0.
5.0 - 5.99	0.	0.	0.	0.	0.
6.0 - 6.99	0.	0.	0.	0.	0.
7.0 - 7.99	0.	0.	0.	0.	0.
8.0 - 8.99	0.	0.	0.	0.	0.
9.0 - 9.99	0.	0.	0.	0.	0.
10. - UP	0.	0.	0.	0.	0.
TOTAL	453.	397.	56.	0.88	0.12

	MEAN	S.D.
CORRECT DOCUMENTS	1.00	0.69
INCORRECT DOCUMENTS	0.57	0.58

```
TABLE 5   DOCUMENT CLASSIFICATION SUMMARY
        140 DOCUMENTS IN EACH CATEGORY   DISCRIMINATING WORD SET    7      48 DISCRIMINATING  WORDS

  TEST
                 AUTO   CATEGORY
ACTUAL
CATEGORY        P2       P3       P4       TOTAL

P2             161.00    10.00    15.00   186.00

P3               8.00   116.00     2.00   126.00

P4              19.00     2.00   120.00   141.00

TOTAL          188.00   128.00   137.00   453.00

                    PERCENTAGE

P2               0.87     0.05     0.08     1.00

P3               0.06     0.92     0.02     1.00

P4               0.13     0.01     0.85     1.00

                         SWETS MEASURES              RECALL        RELEVANCE       PRECISION
CATEGORY P2           PERTINENT   NOT PERTINENT      RATIO           RATIO           RATIO
 RETRIEVED              0.87           0.05           0.87           0.86            1.01


                         SWETS MLASURES              RFCALL        RELEVANCE       PRECISION
CATEGORY P3           PERTINENT   NOT PERTINENT      RATIO           RATIO           RATIO
 RETRIEVED              0.92           0.02           0.92           0.91            1.02


                         SWETS MEASURES              RECALL        RELEVANCE       PRECISION
CATEGORY P4           PERTINENT   NOT PERTINENT      RATIO           RATIO           RATIO
 RETRIEVED              0.85           0.03           0.85           0.88            0.97
```

Table 6. Category Centroids and Dispersions

a. Upper Level

Sample		A		M		P	
Size		I	II	I	II	I	II
35	Centroid	-0.56	0.10	0.46	0.55	0.37	-0.48
	Dispersion	0.10	0.0	0.04	0.02	0.04	-0.03
		0.0	0.03	0.02	0.09	-0.03	0.14
70	Centroid	0.93	0.21	-0.36	-0.35	-0.36	0.76
	Dispersion	0.27	0.03	0.10	0.02	0.12	-0.05
		0.03	0.09	0.02	0.32	-0.05	0.41
140	Centroid	-1.10	0.03	0.61	0.65	0.41	-0.88
	Dispersion	0.57	-0.08	0.19	0.17	0.17	-0.09
		-0.08	0.19	0.17	0.42	-0.09	0.63

b. Lower Level

Sample		A		M		P	
Size		I	II	I	II	I	II
35	Centroid	-0.15	-0.37	0.58	-0.07	-0.08	0.33
	Dispersion	0.02	0.01	0.03	0.0	0.01	-0.01
		0.01	0.03	0.0	0.02	-0.01	0.06
70	Centroid	-0.12	0.67	0.87	-0.24	-0.34	-0.72
	Dispersion	0.03	-0.03	0.14	-0.03	0.07	0.06
		-0.03	0.21	-0.03	0.05	0.06	0.24
140	Centroid	0.15	-0.62	-1.24	0.26	0.50	0.68
	Dispersion	0.14	-0.09	0.44	-0.07	0.14	0.16
		-0.09	0.31	-0.07	0.09	0.16	0.40

Table 7. Sample Error and χ^2 Among Category Centroids

a. Upper Level

	35			70			140		
	A	M	P	A	M	P	A	M	P
Sample Error	1.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0
	0.0	0.99	0.01	0.02	0.83	0.15	0.01	0.93	0.06
	0.0	0.0	1.0	0.01	0.14	0.85	0.02	0.04	0.93
χ^2	0.0	24.1	25.0	0.0	16.1	14.7	0.0	18.8	13.4
	17.4	0.0	10.4	8.1	0.0	3.1	9.3	0.0	4.8
	22.2	12.2	0.0	11.4	3.9	0.0	6.6	7.1	0.0

b. Lower Level

	35			70			140		
	P2	P3	P4	P2	P3	P4	P2	P3	P4
Sample Error	1.0	0.0	0.0	0.99	0.0	0.01	0.91	0.0	0.09
	0.0	1.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0
	0.0	0.0	1.0	0.0	0.0	1.0	0.01	0.0	0.99
χ^2	0.0	24.9	10.6	0.0	16.8	8.3	0.0	10.1	4.8
	33.0	0.0	56.0	28.6	0.0	23.7	14.1	0.0	32.4
	16.1	23.7	0.0	15.5	23.1	0.0	10.9	13.3	0.0

Table 8. Distance of Category Centroid from Origin

a. Upper Level

	35			70			140		
	A	M	P	A	M	P	A	M	P
χ^2	3.72	6.01	4.36	3.36	1.48	2.07	2.21	2.02	1.76
Relative Probability	0.54	0.15	0.31	0.22	0.48	0.30	0.26	0.41	0.33

b. Lower Level

	35			70			140		
	A	M	P	A	M	P	A	M	P
χ^2	4.49	11.71	1.99	2.16	5.36	2.57	1.29	3.53	1.80
Relative Probability	0.22	0.01	0.77	0.57	0.11	0.32	0.47	0.15	0.38

REFERENCES

1. Williams, J. H., Statistical Analysis and Classification of Documents, IRAD Task No. 0274, IBM, FSD, Rockville, Maryland, 1963.

2. Williams, J. H., Results of Classifying Documents with Multiple Discriminant Functions, a paper presented at the National Bureau of Standard's Symposium on Statistical Association Methods for Mechanized Documentation, Washington, D. C., April 1964.

3. Hodges, Joseph L. Jr., Discriminatory Analysis: 1. Survey of Discriminatory Analysis, Report Number 1, Project 21-49-004, USAF School of Aviation Medicine, Randolph Field, Texas, October 1950, ATI-99384.

4. Tatsuoka, Maurice M., and Tiedeman, David V., "Discriminant Analysis," Review of Educational Research, Vol. 24, No. 5, pp. 402-416.

5. Fisher, R. A., "The Use of Multiple Measurements in Taxonomic Problems," Ann. Eugenics, Vol. 7 (1936), pp. 179-188.

6. Barnard, M. M., "The Secular Variations of Skull Characters in Four Series of Egyptian Skulls," Ann. Eugenics, Vol. 6 (1935), pp. 352-371.

7. Rao, C. Radhakrishna., Advanced Statistical Methods in Biometric Research, New York, Wiley & Sons, 1952.

8. Bryan, J. G., "The Generalized Discriminant Function: Mathematical Foundations and Computational Routine, "Harvard Ed. Review, pp. 90-95, 1951.

9. Steel, Robert G., and Torrie, James H., Principles and Procedures of Statistics, New York, McGraw Hill, 1960.

10. Kullback, Solomon, Information Theory and Statistics, New York, Wiley & Sons, 1959.

11. Hoel, Paul G., Introduction to Mathematical Statistics, New York, Wiley & Sons, 1954.

12. Cooley, William W., and Lohnes, Paul R., Multivariate Procedures for the Behavioral Sciences, New York, Wiley & Sons, 1962.

13. Swets, John A., "Information Retrieval Systems, "Science, Vol. 141, July 19, 1963, pp. 245-250.

UNCLASSIFIED
Security Classification

DOCUMENT CONTROL DATA - R&D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1 ORIGINATING ACTIVITY *(Corporate author)* IBM-FSD 7220 Wisconsin Ave Bethesda, Md 20014	2a REPORT SECURITY CLASSIFICATION Unclassified 2b GROUP

3 REPORT TITLE
Discriminant Analysis for Content Classification

4 DESCRIPTIVE NOTES *(Type of report and inclusive dates)*
Technical Report.(Final)

5 AUTHOR(S) *(Last name, first name, initial)*
Williams, John H., Jr.

6. REPORT DATE February 1966	7a TOTAL NO. OF PAGES 272	7b. NO OF REFS 13
8a CONTRACT OR GRANT NO. AF30(602)-3563 b. PROJECT NO. 4594 c d	9a ORIGINATOR'S REPORT NUMBER(S) 9b OTHER REPORT NO(S) *(Any other numbers that may be assigned this report)* RADC-TR-66-6	

10 AVAILABILITY/LIMITATION NOTICES

Distribution of this document is unlimited.

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY RADC, GAFB, NY 13440

13 ABSTRACT

A series of experiments was performed to investigate the effectiveness and utility of automatically classifying documents through the use of multiple discriminant functions. Classification is accomplished by computing the distance from the mean vector of each category to the vector of observed frequencies of a document and assigning the document to the category having the highest probability. Data concerning the effect of the principal classification parameters on classification performance is reported, based on a data base of approximately 2700 abstracts from the solid state physics field. The parameters studied were the number of sample documents required to define a category, the length of documents, the inter-relationship of the number of sample documents and their lengths, the relation of the number of word types in a document to the number of categories, and performance measures. A higher performance level was obtained when samples of 140 documents were used to define eacn category than with samples of 35 and 70 documents. Classification results obtained on independent test sets of documents ranged from 73 to 92 per cent. The test sets contained 419 and 1333 documents. Results are also reported in terms of Swets' effectiveness measure and Cleverdon's ratios of relevance, recall and precision.

UNCLASSIFIED
Security Classification

UNCLASSIFIED

Security Classification

14 KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Information Storage & Retrieval						
Document Classification						
Content Analysis						
Discriminant Analysis						
Probabilistic Indexing						

INSTRUCTIONS

1. ORIGINATING ACTIVITY: Enter the name and address of the contractor, subcontractor, grantee, Department of Defense activity or other organization *(corporate author)* issuing the report.

2*a*. REPORT SECURITY CLASSIFICATION: Enter the overall security classification of the report. Indicate whether "Restricted Data" is included. Marking is to be in accordance with appropriate security regulations.

2*b*. GROUP: Automatic downgrading is specified in DoD Directive 5200 ^ and Armed Forces Industrial Manual. Enter the group number. Also, when applicable, show that optional markings have been used for Group 3 and Group 4 as authorized

3. REPORT TITLE: Enter the complete report title in all capital letters. Titles in all cases should be unclassified. If a meaningful title cannot be selected without classification, show title classification in all capitals in parenthesis immediately following the title.

4. DESCRIPTIVE NOTES: If appropriate, enter the type of report, e.g., interim, progress, summary, annual, or final. Give the inclusive dates when a specific reporting period is covered.

5. AUTHOR(S): Enter the name(s) of author(s) as shown on or in the report. Enter last name, first name, middle initial. If military, show rank and branch of service. The name of the principal author is an absolute minimum requirement.

6. REPORT DATE: Enter the date of the report as day, month, year; or month, year. If more than one date appears on the report, use date of publication.

7*a*. TOTAL NUMBER OF PAGES: The total page count should follow normal pagination procedures, i.e., enter the number of pages containing information.

7*b*. NUMBER OF REFERENCES: Enter the total number of references cited in the report.

8*a*. CONTRACT OR GRANT NUMBER: If appropriate, enter the applicable number of the contract or grant under which the report was written.

8*b*, 8*c*, & 8*d*. PROJECT NUMBER: Enter the appropriate military department identification, such as project number, subproject number, system numbers, task number, etc.

9*a*. ORIGINATOR'S REPORT NUMBER(S): Enter the official report number by which the document will be identified and controlled by the originating activity. This number must be unique to this report.

9*b*. OTHER REPORT NUMBER(S): If the report has been assigned any other report numbers *(either by the originator or by the sponsor)*, also enter this number(s).

10. AVAILABILITY/LIMITATION NOTICES: Enter any limitations on further dissemination of the report, other than those imposed by security classification, using standard statements such as:

(1) "Qualified requesters may obtain copies of this report from DDC."

(2) "Foreign announcement and dissemination of this report by DDC is not authorized."

(3) "U. S. Government agencies may obtain copies of this report directly from DDC. Other qualified DDC users shall request through

_______________________________________."

(4) "U. S. military agencies may obtain copies of this report directly from DDC. Other qualified users shall request through'

_______________________________________."

(5) "All distribution of this report is controlled. Qualified DDC users shall request through

_______________________________________."

If the report has been furnished to the Office of Technical Services, Department of Commerce, for sale to the public, indicate this fact and enter the price, if known.

11. SUPPLEMENTARY NOTES: Use for additional explanatory notes.

12. SPONSORING MILITARY ACTIVITY: Enter the name of the departmental project office or laboratory sponsoring *(paying for)* the research and development. Include address.

13 ABSTRACT: Enter an abstract giving a brief and factual summary of the document indicative of the report, even though it may also appear elsewhere in the body of the technical report. If additional space is required, a continuation sheet shall be attached.

It is highly desirable that the abstract of classified reports be unclassified. Each paragraph of the abstract shall end with an indication of the military security classification of the information in the paragraph. represented as *(TS)*, *(S)*, *(C)*, or *(U)*

There is no limitation on the length of the abstract. However, the suggested length is from 150 to 225 words.

14. KEY WORDS. Key words are technically meaningful terms or short phrases that characterize a report and may be used as index entries for cataloging the report. Key words must be selected so that no security classification is required. Identifiers, such as equipment model designation, trade name, military project code name, geographic location, may be used as key words but will be followed by an indication of technical context. The assignment of links, rules, and weights is optional.

UNCLASSIFIED

Security Classification